IN THE UNITED STATES DISTRICT COURT
FOR THE CENTRAL DISTRICT OF CALIFORNIA

BERNADINE GRIFFITH, et al., individually and on behalf of all others similarly situated, Plaintiffs, v. TIKTOK, INC., a corporation; BYTEDANCE, INC., a corporation, Defendants.	Case No. 5:23-cv-00964-SB-E

EXPERT REPORT OF ZUBAIR SHAFIQ, PH.D.

September 20, 2024

TABLE OF CONTENTS

TABLE OF FIGURES

List of Appendices to Expert Report of Zubair Shafiq, Ph.D. (September 20, 2024)

O	URL Classification of 10,000 Random URL Samples from Produced March 28 and May 21 Processed Unmatched Pixel Data
P	Reversing Hashed Phone Numbers in Produced March 28 and May 21 Processed Unmatched Pixel Data
Q	Script for Extracting Unique URLs in Produced March 28 and May 21 Processed Unmatched Pixel Data
R	URLs in Plaintiff Internet Artifacts and History Data Found in Produced March 28 and May 21 Processed Unmatched Pixel Data
S	"Vignette" Analysis of TikTok's Collection of Data on One Non-TikTok User on March 28, 2024 and May 21, 2024
T	Materials Reviewed and Relied Upon

I. QUALIFICATIONS AND ASSIGNMENT

1. My name is Zubair Shafiq, Ph.D. I am an Associate Professor of Computer Science at the University of California-Davis, where I lead a research lab focused on online privacy, security, and safety.

2. My lab's research aims to uncover personal data collection, sharing, and usage in the online advertising ecosystem.

3. In addition to my research, I regularly teach undergraduate and graduate courses on computer networks and computer security, including special topics courses covering emerging trends in online advertising and tracking.

4. My research is funded by the National Science Foundation (NSF) through multiple highly competitive research grants. Notably, I am leading the National Science Foundation (NSF) Secure and Trustworthy Cyberspace (SaTC) Frontier Center on Protecting Personal Data Flow on the Internet (ProperData). As part of this effort, my research group is building new device instrumentation systems and measurement methods to investigate personal data collection, sharing, and usage in the web, mobile, and Internet-of-Things (IoT) ecosystems.

5. I have received several awards and distinctions for my research. I am a recipient of the Caspar Bowden Award - Runner-up for Outstanding Research in Privacy Enhancing Technologies (2024), Chancellor's Fellowship (2022-2023), Dean's Scholar Award (2020), National Science Foundation CAREER Award (2018), and Fitch-Beach Outstanding Graduate Research Award (2013).

6. I have co-authored more than 100 peer-reviewed research papers. I received the Best Paper Award at the 2023 ACM Internet Measurement Conference for my research on tracking, profiling, and ad targeting in the Amazon Alexa ecosystem. I also received the 2018 Andreas

Pfitzmann Award at the Privacy Enhancing Technologies Symposium for my research on designing a system to detect advertising and tracking data flows in mobile apps. I also received the Best Paper Award at the 2017 ACM Internet Measurement Conference for my research on identifying and investigating the abuse of a security vulnerability in Facebook Graph API's implementation of third-party apps. I also received the Best Paper Award at the 2012 IEEE International Conference on Network Protocols for my research on reverse-engineering proprietary network protocols through network traffic analysis.

7. I am the editor-in-chief of the Proceedings on Privacy Enhancing Technologies (PoPETs). I am on the steering committee of the Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb). I am the general chair of the Workshop on Technology and Consumer Protection (ConPro). In the past, I have served as the program chair for the Workshop on Technology and Consumer Protection (ConPro 2022 and 2023) and the Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb 2022 and 2023).

8. My complete CV is attached as **Appendix A**.

9. I have been retained by counsel for Plaintiffs to serve as an independent expert in this litigation.

10. On June 21, 2024, I submitted the Declaration of Zubair Shafiq, Ph. D., in Support of Plaintiffs' Motion for Class Certification ("Shafiq Opening Declaration"). On July 26, 2024, I submitted the Reply Declaration of Zubair Shafiq, Ph.D., in Support of Plaintiffs' Motion for Class Certification ("Shafiq Reply Declaration"). I incorporate by reference both the Shafiq Opening Declaration and the Shafiq Reply Declaration into this Report.

11. I am compensated at the rate of $750/hour. My compensation is not dependent on and in no way affects the substance of my opinions. Nor does my compensation depend on the

outcome of this proceeding. I understand that, should there be any recovery in this case, I will be excluded from any disbursement of funds.

12. Materials reviewed and relied upon for this report are identified in the attached **Appendix T**.

13. I reserve the right to amend, modify and supplement this report should new or additional information be made available to me.

II. SUMMARY OF THE OPINIONS PROFFERED

14. In addition to the opinions offered in the Shafiq Opening Declaration and Shafiq Reply Declaration, I offer the following opinions and conclusions, which are consistent with those offered in the Opening and Reply Declarations.

15. Opinion No. 1: Based on my review of the March 28, 2024 and May 21, 2024 one-day sample data produced by TikTok, TikTok has collected data [redacted]

16. Opinion No. 2: My review of the plaintiffs' browsing history of websites that are included in TikTok's two one-day sample data sets further corroborates that TikTok collected data [redacted]

[redacted]

17. Opinion No. 3: Hashing an email address or phone number, without additional security measures, does not provide meaningful privacy protection. Indeed, my analysis throughout this report confirms that a hashed email address or phone number can be trivially reversed such that the data collected by TikTok on non-TikTok users can readily be associated with a plaintext email address or phone number.

18. Opinion No. 4: [redacted] My entropy analysis shows that IP address, user agent, and cookies collected by TikTok together easily exceed the 32-bit entropy threshold, which the industry sets as the threshold above which bits of entropy are "enough to uniquely identify every individual person."

19. Opinion No. 5: Consistent with TikTok's collection of data on the named plaintiffs, TikTok collects ***sensitive*** data on other non-TikTok users. [redacted] This vignette demonstrates just how much TikTok can glean about a single non-TikTok user from the data it collects in a short time period.

20. Opinion No. 6: The data that TikTok collected from the named plaintiffs is typical of the data TikTok collected from non-TikTok users at large. I reach this conclusion based on two findings: [redacted]

21. Opinion No. 7: Preliminary analysis of source code confirms the opinion in the Shafiq Opening Declaration that [redacted] I reserve the right to amend, modify, and/or supplement my opinions based on the production of source code and documents that were not available to me as I was preparing this report.

22. Opinion No. 8: It is possible to write and run a computer program that crawls privacy policies of websites and programmatically detects whether the privacy policies mention the word "TikTok" or "ByteDance." For this report, I programmatically searched for "TikTok" and "ByteDance" across the privacy policies of websites that accounted for 54% of the event data in the two one-day sample data sets produced by TikTok. Only 7.5% of these websites mentioned TikTok; none mentioned ByteDance, with the exception of one website which stated that ByteDance is affiliated with TikTok. Only 2.5% mentioned some type of data collection by TikTok. None of the websites describe the full extent of data collection by TikTok Pixel and Events API.

23. I elaborate on each of these opinions below and accompanying appendices.

III. BACKGROUND OF RELEVANT TECHNOLOGY

A. Overview of Pixels

24. The term pixel (also known as tracking pixel, web bug, pixel tag, or web beacon) refers to a piece of code (e.g., JavaScript[1] or HTML[2]) or image (e.g., a 1x1 GIF[3]) that is used to track browsing activity on the web.

25. A pixel typically allows a third party[4] (i.e., a domain[5] or origin[6] that is distinct from the first-party website that a website visitor navigates to) to track a website visitor visiting website$_1$ at time$_1$, website$_2$ at time$_2$, and so on. Pixels can track more fine-grained activity on and across websites, such as full webpage URLs,[7] webpage title, search terms, forms fields, adding an item to cart, etc.

26. [redacted]

[redacted]

[1] https://developer.mozilla.org/en-US/docs/Web/JavaScript.

[2] https://developer.mozilla.org/en-US/docs/Web/HTML.

[3] Ruohonen, J. and Leppänen, V., 2018, January. *Invisible pixels are dead, long live invisible pixels!*, in Proceedings of the 2018 Workshop on Privacy in the Electronic Society (pp. 28-32).

[4] https://web.dev/learn/privacy/third-parties.

[5] https://developer.mozilla.org/en-US/docs/Glossary/Domain.

[6] https://developer.mozilla.org/en-US/docs/Glossary/Origin.

[7] https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Web_mechanics/What_is_a_URL.

[redacted][8]

27. A pixel is designed to collect two types of data in HTTP[9] request[10] transmissions from a website visitor's web browser to the pixel's web server: (1) identifiers and (2) browsing activity.

a. Identifiers are typically collected via (i) the cookies[11] stored in the web browser's storage[12] and (ii) the combination of IP address[13] and user agent[14] in the transmission from a website visitor's web browser to the pixel's web server.[15] Cookies containing identifiers typically store a 128 bit Universally Unique IDentifier (UUID)[16] that is sufficient for unique identification. There are two main types of cookies: first-party and third-party. First-party cookies are set on the same domain as the visited website's domain. Third-party cookies are set on a different domain as the visited website's domain. Pixels set third-party cookies that allow them to track users across websites. Since some web browsers have now started to restrict third-party cookies, pixels now also set first-party cookies[17] (a practice

[8] [redacted]

[9] https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview.

[10] https://developer.mozilla.org/en-US/docs/Web/HTTP/Messages#http_requests.

[11] https://developer.mozilla.org/en-US/docs/Web/HTTP/Cookies.

[12] Web browsers support several cookie-like storage mechanisms such as cookies, session storage, local storage, cache storage, and IndexedDB. See https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Client-side_web_APIs/Client-side_storage for more details.

[13] https://developer.mozilla.org/en-US/docs/Glossary/IP_Address.

[14] https://developer.mozilla.org/en-US/docs/Glossary/User_agent.

[15] IAB Tech Lab Identity Solutions Guidance Version 1.0 https://iabtechlab.com/wp-content/uploads/2024/05/Identity-Solutions-Guidance-FINAL.pdf.

[16] https://datatracker.ietf.org/doc/html/rfc4122.

[17] Cookies or their equivalent storage mechanisms such as local storage or session storage.

known as cookie ghostwriting[18,19]) on the same domain as the visited website's domain to circumvent third-party cookie blocking.[20] First-party cookies are used by pixels for both same-site and cross-site tracking.[21,22] The combination of IP address and user agent typically contains sufficiently distinguishing information[23] (quantified in terms of entropy bits[24,25,26,27]) to be used as a unique identifier.[28] This practice of combining IP address, user agent, and other distinguishing browser or device information for identification is also known as fingerprinting.[29]

[18] Sanchez-Rola, I., Dell'Amico, M., Balzarotti, D., Vervier, P.A. and Bilge, L., 2021, May. Journey to the center of the cookie ecosystem: Unraveling actors' roles and relationships. In 2021 IEEE Symposium on Security and Privacy (SP) (pp. 1990-2004). IEEE.

[19] Nikkhah Bahrami, P., Fass, A. and Shafiq, Z., 2024. COOKIEGUARD: Characterizing and Isolating the First-Party Cookie Jar. arXiv e-prints, pp.arXiv-2406.

[20] Munir, S., Siby, S., Iqbal, U., Englehardt, S., Shafiq, Z. and Troncoso, C., 2023, November. CookieGraph: Understanding and Detecting First-Party Tracking Cookies. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (pp. 3490-3504).

[21] Munir, S., Lee, P., Iqbal, U., Shafiq, Z. and Siby, S., 2024. PURL: Safe and Effective Sanitization of Link Decoration. In 33rd USENIX Security Symposium (USENIX Security 24) (pp. 4103-4120).

[22] Bekos, P., Papadopoulos, P., Markatos, E.P. and Kourtellis, N., 2023, April. The Hitchhiker's guide to facebook web tracking with invisible pixels and click IDs. In Proceedings of the ACM Web Conference 2023 (pp. 2132-2143).

[23] https://clearcode.cc/blog/adtech-id-solutions/ ("A universal ID is a unique user ID that allows AdTech companies to identify users across different websites and devices. Universal IDs are created using **probabilistic data** (e.g. IP address, browser type and model, and user-agent string) or **deterministic data** (e.g. an email address or phone number), or both, to produce an ID." (emphasis in original)).

[24] Eckersley, P. (2010). How unique is your web browser?. In Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010.

[25] Wagner, I. and Eckhoff, D., 2018. Technical privacy metrics: a systematic survey. ACM Computing Surveys, 51(3).

[26] Google. Introducing the Privacy Budget, https://www.youtube.com/watch?v=0STgfjSA6T8&t=423s.

[27] https://github.com/mikewest/privacy-budget/blob/4e5f78adde92bd622dafeceae78682fc0823c0eb/faq.md.

[28] IP address also encodes information about the location of a user. There are numerous IP geolocation services that can estimate the country, state, city, postal code, and even approximate longitude and latitude from IP address. For example, see https://www.maxmind.com/en/geoip-web-services-demo.

[29] Yen, T.F., Xie, Y., Yu, F., Yu, R.P. and Abadi, M., 2012, February. Host Fingerprinting and Tracking on the Web: Privacy and Security Implications. In Network and Distributed System Security (NDSS) Symposium.

b. Browsing activity is collected via the (i) URL,[30] (ii) Referer[31] and Origin[32] headers, or (iii) payload[33] of the transmission from a website visitor's web browser to the pixel's web server. The URL may contain information about the webpage's content (e.g., name of a product) in a query parameter.[34,35] The Referer and Origin headers typically contain the name of the website visited by the website visitor. The payload may contain much more detailed content information[36] such as the detailed description of a product in a standardized format such as JSON,[37] which is a widely used way to store data in a standard human-readable text format that is also amenable to automated machine parsing.

28. The data collected by a pixel is typically used to target website visitors with ads that are personalized based on their browsing history. For example, a website visitor whose browsing history indicates interest in hiking may receive targeted ads for hiking poles.[38] After a personalized ad is served, pixels are also used to collect information about whether a website visitor viewed or clicked on an ad and ended up buying the advertised product.[39] This information is then used to further personalize ads. For example, a website visitor may get a personalized ad

[30] https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Web_mechanics/What_is_a_URL.

[31] https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Referer.

[32] https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Origin.

[33] https://developer.mozilla.org/en-US/docs/Glossary/Payload_body.

[34] https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Web_mechanics/What_is_a_URL#parameters.

[35] URL parameters may also contain identifiers such as cookies, email address, or phone number.

[36] Payload may also contain identifiers such as cookies, email address, or phone number.

[37] https://developer.mozilla.org/en-US/docs/Glossary/JSON.

[38] https://clearcode.cc/glossary/ad-targeting/.

[39] https://clearcode.cc/glossary/conversion-pixel/.

for a hiking pole that they saw in an ad and added to cart but did not buy yet. In addition, the data collected by a pixel can be used to improve various related systems such as fraud detection[40] and user identification and targeting algorithms.[41]

B. Overview of TikTok Pixel

29. TikTok Pixel is JavaScript source code that is embedded on non-TikTok websites. TikTok also describes the TikTok Pixel as: “TikTok Pixel is a piece of code that you can place on your website that allows you to share website events with TikTok.”[42]

30. TikTok Pixel’s source code is written by TikTok and is served by TikTok’s server. The website developers of the non-TikTok websites[43] that use the Pixel do not write the source code, nor can they directly modify it.

31. [redacted][redacted] [redacted][redacted] TikTok Pixel’s placement in the webpage header also ensures that its data collection is real-time and contemporaneous with the loading of the webpage.

[40] https://clearcode.cc/blog/rtb-online-advertising-fraud/.

[41] IAB Tech Lab. Identity Solutions Guidance Version 1.0, https://iabtechlab.com/wp-content/uploads/2024/05/Identity-Solutions-Guidance-FINAL.pdf.

[42] https://ads.tiktok.com/help/article/tiktok-pixel.

[43] Because websites that use the TikTok Pixel typically advertise their products and services on the TikTok app or web platform, TikTok’s internal documents commonly refer to these non-TikTok websites as “advertisers” or “clients.”

[44] [redacted]

[45] https://www.corewebvitals.io/pagespeed/head-vs-footer-javascript-and-core-web-vitals.

32. According to TikTok's own public documentation,[46, 47] the data collected by TikTok Pixel includes at least the following: "Timestamp,"[48] "Cookies,"[49] "IP Address,"[50] "User Agent,"[51] "Ad/Event information,"[52] and "Metadata & Button Clicks."[53,54]

33. TikTok Pixel collects this data through "standard events" and "custom events":

a. Standard events include "Add Payment Info," "Add to Cart," "Add to Wishlist," "Click Button," "Complete Payment," "Complete Registration," "Contact," "Download," "Initiate Checkout," "Place an Order," "Search," "Submit Form," "Subscribe," and "View Content."[55]

b. Custom events are "actions that TikTok partners can define themselves beyond the predefined standard events list."[56]

[46] https://ads.tiktok.com/help/article/tiktok-pixel.

[47] [redacted] TIKTOK-BG-000157260, at -262; *see also* TIKTOK-BG-8008 at -010 [redacted]

[48] https://ads.tiktok.com/help/article/tiktok-pixel ("Used to determine when website actions took place, like when a page was viewed or when a product was purchased").

[49] *Id.* ("Used to help with the measurement, optimization, and targeting of your campaigns. First-party cookies are optional, but third-party cookies are on by default with the TikTok Pixel. Performance is boosted when first- and third-party cookies are paired with Advanced matching").

[50] *Id.* ("Used to determine the geographic location of an event").

[51] *Id.* ("Used to determine the device make, model, operating system, and browser information").

[52] *Id.* ("Information about the ad a person on TikTok has clicked on or an event that was initiated").

[53] *Id.* ("Includes descriptive page metadata, structured microdata, page performance data, and button clicks").

[54] As discussed below, TikTok Pixel also automatically collects Page URL and Referrer URL by default. Notably, URL is missing from TikTok's public documentation about what data TikTok Pixel collects.

[55] https://ads.tiktok.com/help/article/supported-standard-events; *see also* TIKTOK-BG-000000128.

[56] TIKTOK-BG-000000128.

34. In addition to these standard and custom events, TikTok, by default, collects data through the "PageView" event, [redacted][57] TikTok Pixel, [redacted] [redacted]

35. The scale of TikTok Pixel's data collection is staggering. According to multiple independent estimates, TikTok Pixel is used by more than 300 thousand websites.[60,61] An April 2023 study on the prevalence of pixels have estimated that the TikTok Pixel is present on 7.41 percent of over 3,100 websites spanning a range of industries, including Financial Services & Banking, Healthcare, Technology and SaaS, e-Commerce, Airlines, and U.S. Federal and State

[57] TIKTOK-BG-000000875 (Depo Ex. 55) at -878 ("[redacted]

[58] TIKTOK-BG-000000128; *see also* https://web.archive.org/web/20231129091101/https://ads.tiktok.com/help/article/standard-events-parameters?lang=en. At some point after the commencement of this litigation, TikTok removed this disclosure of the default collection of PageView event data from its website.

[59] *See, e.g.*, TIKTOK-BG-000151364 (Depo Ex. 46), at -366 [redacted] *id.* at -367 [redacted] *id.* at -377 [redacted]; TIKTOK-BG-000151574 (Depo Ex. 47), at -576 [redacted] TIKTOK-BG-000150667 (Depo Ex. 49) ("[redacted] *id.* at -669 [redacted] Tr. of Becca Wong Depo. (May 17, 2024) at 133:1-5 [redacted]

[60] https://trends.builtwith.com/websitelist/TikTok-Conversion-Tracking-Pixel/United-States.

[61] https://www.nerdydata.com/reports/tiktok-pixel/de68a0d2-1056-47f0-aec4-6f705982fc81.

Government.[62] Another study dated March 2024 reports that the TikTok Pixel is present on 12 percent of 3,419 websites spanning healthcare, technology, financial services, retail, and sites of companies listed in the S&P 500 index.[63] [redacted][65]

36. [redacted][67] [redacted][69] [redacted]

[62] Feroot, "Beware of Pixels & Trackers," 2023 Feroot Client-Side Security Report, Apr. 5, 2023. https://go.feroot.com/hubfs/4605309/Reports/Beware%20of%20Pixels%20&%20Trackers%20-%20Feroot%20Client-Side%20Security%20Report%20March%202023.pdf.

[63] LOKKER, "Website Privacy and Compliance Challenges: Quantifying Website Privacy Risks," Mar. 2024, https://lokker.com/wp-content/uploads/2024/04/LOKKER_Online-Data-Privacy-Report_032024-2.pdf.

[64] TIKTOK-BG-000003014, at -3016.

[65] TIKTOK-BG-000009897 at -897. [redacted] TIKTOK-BG-000157229 at -230.

[66] TIKTOK-BG-000086213 at -214.

[67] TIKTOK-BG-000009897 at -897.

[68] TIKTOK-BG-000005393 at -394 [redacted].

[69] *Id.*

C. Overview of TikTok Events API

37. [redacted] The main difference between TikTok Events API and TikTok Pixel is that the former is a "server-to-server" (between website's and TikTok's servers) data-collection mechanism and the latter is "client-to-server" (between website visitor's web browser and TikTok's server) data-collection mechanism.

38. [redacted] and tracking protection features.[72,73,74]

39. TikTok recommends "setting up both TikTok's Events API and Pixel together."[75,76] Therefore, the data collected by them is often duplicative, which is then deduplicated by TikTok.[77]

[70] [redacted]

[redacted]

[72] https://stape.io/blog/how-to-set-up-tiktok-events-api ("With server-side tracking, you will be able to collect more events. TikTok events API is resistant to ad blockers, ITPs, and other tracking restrictions.").

[73] https://ads.tiktok.com/help/article/events-api ("**Resilient solutions for an evolving advertising ecosystem:** The Events API together with an existing Pixel ensures a more sustainable transition in response to changes in the advertising industry.").

[74] https://www.tiktok.com/business/en-US/blog/events-api-consolidated-endpoint ("TikTok launches enhanced Events API with consolidated endpoint"); *id.* ("The era of third-party cookies as we know it is ending. Internet users are looking for more control over their data and how it is used. The use of ad blockers and secure web browsers are on the rise as a result. At the same time, US state governments are signing into law new regulations and policies protecting user data and increasing requirements for collecting, sharing and using data (CPRA, CTDPA, VCDPA, etc). Finally, operating systems and browsers are implementing technical and policy changes limiting how data is collected and used. This combination of factors is driving the end of third-party cookies." "To help advertisers better prepare for this cookieless future, we're excited to announce the launch of a consolidated endpoint across Events API for Web, App (in testing), and Offline.").

[75] TIKTOK-BG-000001355 at -355.

[76] https://ads.tiktok.com/help/article/events-api ("we recommend having an Events API integration with your existing Pixel integration to maximize performance benefits"); *id.* ("Use both TikTok's Pixel and Events API together").

[77] https://ads.tiktok.com/help/article/event-deduplication.

40. Just like TikTok Pixel, the data collection by the TikTok Events API is real-time and contemporaneous with the loading of the webpage. Specifically, TikTok recommends sending "**the event in real-time**" and "as soon as it is seen on the advertiser's server" when using TikTok Events API.[78] Thus, even though the data collected by TikTok Events API first goes from the website visitor's web browser to the website's server before reaching TikTok's server, it remains real-time and contemporaneous with the loading of the webpage just like the data collected by TikTok Pixel.

41. Just like TikTok Pixel, TikTok Events API can collect identifiers in cookies as well as IP address and user agent.[79]

42. [redacted] [redacted]

43. Just like TikTok Pixel, the data collected by TikTok Events API is used for advertising and other related systems.[81,82]

[78] https://business-api.tiktok.com/portal/docs?rid=p41a33fdhon&id=1771100865818625 ("it's **highly recommended to send the event in real-time (without batching)** as soon as it is seen on the advertiser's server") (emphasis in original).

[79] https://ads.tiktok.com/help/article/how-to-set-up-matching-events-with-events-api (explaining that Events API can be set up to collect and transmit "Click ID," i.e. "[a] unique identifier appended to the URL every time a person clicks on a TikTok ad"; "Email and Phone (Hashing required)"; "External ID (Hashing required)," including "[a]dvertiser-side identifiers, such as loyalty membership IDs, advertiser customer IDs, and external cookie IDs"; "IP Address (IP) and User Agent (UA)" and "1st Party Cookie (if using with Pixel)").

[80] TIKTOK-BG-000001355 at -357; *see also* TIKTOK-BG-000008008 at -010.

[81] https://ads.tiktok.com/help/article/events-api ("TikTok Events API provides advertisers with a reliable connection between TikTok and advertiser's marketing data, across web, app, and offline (eg. Store, CRM) channels with the ability to customize the information they share with TikTok.").

[82] https://www.tiktok.com/business/en-US/blog/events-api-consolidated-endpoint ("Events API is a secure server-to-server (S2S) integration with TikTok that allows advertisers to share marketing data with us in a secure connection directly from their server. By sharing this marketing data with TikTok, advertisers are able to unlock the performance advertising benefits of better optimization and delivery.").

44. [redacted][redacted]

45. Just like with TikTok Pixel, the data collected by TikTok Events API is agnostic to whether a website visitor is a TikTok user or not. In the same vein, most of the data collected by TikTok Events API is for non-TikTok users. [redacted][redacted] Thus, just as with TikTok Pixel, most—if not the vast majority—of the data collected by TikTok Events API is for non-TikTok users.

IV. TIKTOK'S UNIFORM DATA COLLECTION FROM NON-TIKTOK USERS

46. I investigated TikTok Pixel's data collection from non-TikTok users across different websites in the following two ways:

a. I crawled a random sample of websites from processed data produced by TikTok (TIKTOK-BG-0124043). To this end, I first compiled the list of unique website domains from that document and then randomly sampled a URL associated for each randomly sampled domain. Each URL was crawled using a fresh Chrome browser (version 125) instance in its default setting that was automated using ChromeDriver[86] (configured to simulate a website visitor that simply loads the

[83] TIKTOK-BG-000003172 at -175.

[84] TIKTOK-BG-000003172 at -175.

[85] TIKTOK-BG-000003172 at -176.

[86] https://developer.chrome.com/docs/chromedriver.

webpage and does not engage in any interaction on the webpage such as click on any cookie disclosure or consent banners) and the network traffic logs were collected using ChromeDriver's built-in logging feature.[87] For each sample, I analyzed the network traffic logs to confirm that there are transmissions to TikTok's server (analytics.tiktok.com). This is to ensure that TikTok Pixel is deployed on the URL. This process continued until there were 1,000 URLs, each with a TikTok Pixel transmission. The source code of my crawls and the list of these 1,000 "Random Sample" URLs are provided in **Appendix C**.

b. Separately, I crawled top-ranked websites in a list of websites produced by TikTok (TIKTOK-BG-000002788), identified as having used the TikTok Pixel.[88] To this end, I first compiled the list of unique website domains and then identified the top-ranked domains using the Tranco[89] ranking. For each domain, a randomly selected URL from the processed data produced by TikTok (TIKTOK-BG-0124043) was crawled using a fresh Chrome browser (version 125) instance in its default setting that was automated using ChromeDriver[90] (configured to simulate a website visitor that simply loads the webpage and does not engage in any interaction on the webpage such as click on any cookie disclosure or consent banners) and the network traffic logs were collected using ChromeDriver's built-in logging

[87] https://developer.chrome.com/docs/chromedriver/logging/performance-log.

[88] [redacted]

[89] Pochat, V.L., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M. and Joosen, W., 2018. Tranco: A research-oriented top sites ranking hardened against manipulation. arXiv preprint arXiv:1806.01156.

[90] https://developer.chrome.com/docs/chromedriver.

feature.[91] For each sample, I analyzed the network traffic logs to confirm that there are transmissions to TikTok's server (analytics.tiktok.com). This is again to ensure that TikTok Pixel is deployed on the URL. This process continued until there were 1,000 URLs, each with a TikTok Pixel transmission. The source code of my crawls and the list of these 1,000 "Top-Ranked" URLs are provided in **Appendix C**.

47. The following table reports the percentage of the 1,000 "Random Sample" and 1,000 "Top-Ranked" URLs where each of the seven data categories are collected by TikTok. It is evident that there is no substantial variability in TikTok Pixel's collection of the seven data categories across both (random sample and top-ranked) sets of websites.

	Random Sample	**Top-Ranked**
Timestamp	100.0%	100.0%
IP Address	100.0%	100.0%
User Agent	100.0%	100.0%
Cookies[92]	100.0%	100.0%
URL[93]	100.0%	100.0%
Event Information[94]	100.0%	100.0%
Content Information[95]	98.0%	97.2%

48. The small fraction of websites from which TikTok Pixel does not collect Content Information either do not have the webpage set up using one of the machine-readable formats supported by TikTok Pixel or have toggled off the default Enhance Data Postback.[96] The former case represents a scenario where TikTok Pixel attempted to collect Content Information but was

[91] https://developer.chrome.com/docs/chromedriver/logging/performance-log.

[92] Third-party cookie, First-party cookie, or Session ID.

[93] Page URL or Referrer URL.

[94] Event or Message ID.

[95] content_data or properties.

[96] https://ads.tiktok.com/help/article/enhance-data-postback-with-the-tiktok-pixel.

unable to due to the website's formatting. Thus, the results in the above table are a lower bound of the webpages where TikTok Pixel attempts to collect the seven data categories. The latter case can be automatically detected and excluded from analysis if necessary.[97]

49. I further investigated the variability in TikTok Pixel's data collection across different web browsers in two ways:

a. I crawled six websites[98] included in the Second Amended Complaint using four major web browsers (Chrome, Safari, Edge, Firefox) that, combined, account for more than 95% of the browser market share in the United States.[99] For each of the six websites visited, I navigated to the homepage, conducted a search on that homepage if that was an option, and clicked on one subpage. Other than search or clicking on a subpage, I did not interact with any cookie disclosure or consent banners, if available. Each webpage was crawled using a fresh browser[100] instance in its default setting, and the network traffic logs were collected using each browser's respective developer tools. The underlying data, containing further details about my crawls, is provided in **Appendix C**.

b. Separately, I analyzed the processed data produced by TikTok (TIKTOK-BG-0124043[101]) across the four major web browsers (Chrome, Safari, Edge, Firefox) that together account for more than 95% of the browser market share in the United

[97] A corner case that I found and handled was that TikTok Pixel sent data in a GET request rather than a POST request.

[98] The six websites are buildabear.com, etsy.com, hulu.com, riteaid.com, upwork.com, and vitaminshoppe.com.

[99] https://gs.statcounter.com/browser-market-share/all/united-states-of-america. My testing focuses on the default settings of the web browsers.

[100] Chrome version 125, Safari version 16.1, Edge version 125, Firefox version 126.

[101] My analysis focuses on the rows where "[redacted] (i.e., the data was collected by TikTok Pixel) and [redacted] (the data was not matched to TikTok registered user or anonymous TikTok user).

States.[102] This analysis covers TikTok Pixel's data collection on tens of thousands of data points for a diverse range of web browser configurations across the four major web browsers. The source code of my analysis is provided in **Appendix C**.

50. The following table reports whether the seven data categories are collected by TikTok on riteaid.com, one of the six websites. It is evident that there is no substantial variability in TikTok Pixel's collection of the seven data categories from riteaid.com across the four web browsers. The results are similar for the other six websites – i.e., there is no substantial variability in TikTok Pixel's collection of the seven default data categories across the four web browsers.

	Chrome	Safari	Edge	Firefox
Timestamp	✓	✓	✓	✓
IP Address	✓	✓	✓	✓
User Agent	✓	✓	✓	✓
Cookies[103]	✓	✓	✓	✓
URL[104]	✓	✓	✓	✓
Event Information[105]	✓	✓	✓	✓
Content Information[106]	✓	✓	✓	✓

51. The following table reports the percentage of URLs where each of the seven data categories are collected by TikTok. It is evident that there is no [redacted] in TikTok Pixel's collection of the seven data categories across the four web browsers.

	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]

[102] https://gs.statcounter.com/browser-market-share/all/united-states-of-america.

[103] Third-party cookie, First-party cookie, or Session ID.

[104] Page URL or Referrer URL.

[105] Event or Message ID.

[106] content_data or properties.

[107] In the column "[redacted] of TIKTOK-BG-0124043.

[108] In the column [redacted] of TIKTOK-BG-0124043.

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

V. TIKTOK'S DATA COLLECTION FOR THE NAMED PLAINTIFFS

52. In the Shafiq Opening Declaration, I described how, given the sheer scale of TikTok Pixel's data collection on hundreds of thousands of websites, I conducted statistical analysis to offer the opinion that it is unlikely that there is a non-trivial number of Internet users (which would include non-TikTok users such as the plaintiffs in this action who testified that they visited websites where TikTok collects data[114]) in the United States for whom TikTok Pixel has not collected data at least once during the relevant time period.[115]

53. Given that TikTok has now produced two one-day samples of the processed data for March 28, 2024[116] and May 21, 2024,[117] I use this data to provide further evidence that

[109] In the column [redacted] of TIKTOK-BG-0124043.

[110] Third-party _ttp cookie in the "[redacted] column; First-party _ttp cookie in the "[redacted] column; or Session storage tt_sessionId in the "[redacted] column of TIKTOK-BG-0124043.

[111] Page URL in the "[redacted] column or Referrer URL in the "[redacted] column of TIKTOK-BG-0124043.

[112] Event in the "[redacted] column or Message ID in the "[redacted]" column of TIKTOK-BG-0124043.

[113] The data produced by TikTok (TIKTOK-BG-0124043) from the [redacted]' table does not include all of the content information, such as properties ([redacted] of the webpage content in one of the four standard formats (e.g., JSON-LD, OpenGraph), that TikTok Pixel automatically collects in POST request transmissions to https://analytics.tiktok.com/api/v2/pixel/act.

[114] *See* Declaration of Bernadine Griffith at ¶¶ 4-6 [ECF No. 177-2]; Transcript of Bernadine Griffith Deposition at 62:10-18, 64:22-65:3; 68:16-24; Declaration of Patricia Shih at ¶¶ 4-6 [ECF No. 177-3]; Transcript of Patricia Shih Blough Deposition at 49:13-15; 50:22-51:17; Declaration of Jacob Watters at ¶¶ 4-6 [ECF No. 177-4]; Transcript of Jacob Watters Leady Deposition at 50:23-51:2.

[115] Shafiq Opening Declaration at ¶ 95.

[116] Files: [redacted]

[117] Files: [redacted]

confirms that TikTok indeed collected data from non-TikTok users, including the plaintiffs in this action. It is important to note that even these two one-day samples of the processed data contain only a small fraction of all the data collected by TikTok from non-TikTok users, including the plaintiffs, over the relevant time period.

54. I conduct the following analysis to investigate TikTok's data collection for the named plaintiffs:

a. Evidence of the plaintiffs' data in the two one-day samples of the processed data produced by TikTok for March 28, 2024[118] and May 21, 2024.[119]

b. Evidence of TikTok's data collection in the websites with TikTok Pixel installed that plaintiffs Griffith,[120] Shih,[121] and Watters visited.[122]

A. Evidence of Plaintiffs' Data in the One-Day Data Samples Produced by TikTok

55. Since TikTok collects multiple identifiers such as IP address and user agent, cookies, email address (both plaintext and hashed), and phone number (both plaintext and hashed), I first searched for the identifiers of the named plaintiffs, such as their email addresses and cookies,

[118] Files: [redacted]

[119] Files: [redacted]

[120] Files: [redacted] (produced as GRIFFITHTT002116), [redacted] (produced as GRIFFITHTT002117) and Appendix R Plaintiff Internet Artifacts and History Data.

[121] Files: [redacted] (produced as SHIH-GRIFFITHTT000183), [redacted] (SHIH-GRIFFITHTT000184), [redacted] (SHIH-GRIFFITHTT000185), [redacted] SHIH-GRIFFITHTT000186), and Appendix R Plaintiff Internet Artifacts and History Data.

[122] Files: [redacted] (produced as WATTERS-GRIFFITHTT000610), [redacted] (WATTERS-GRIFFITHTT000611)[redacted] (WATTERS-GRIFFITHTT000612) and Appendix R Plaintiff Internet Artifacts and History Data.

in the two sets of one-day processed data produced by TikTok and then further searched for the identifiers contained therein.

56. It is important to recognize that hashed email address or phone number can be trivially reversed and linked to an individual, for as little as 2 to 4 cents. There are numerous services that allow reversing hashed email addresses or phone numbers as well as append additional identifiers such as name and physical address. For example,

a. As shown in Figure 1, Datafinder reverses hashed email addresses—including the SHA-256 hash used by TikTok—for 4 cents. It further appends the first name, last name, address, city, state, zip, and phone number associated with the hashed email address for an additional 4 cents.[123]

b. As shown in Figure 2, DataZapp also reverses hashed email addresses—including the SHA-256 hash used by TikTok—for 3 cents. For an additional 3 cents, DataZapp adds names, addresses, and phone numbers to the reversed hashed email addresses.[124,125,126,127]

c. As shown in Figure 3, DataGroup also reverses hashed email addresses and phone numbers for 2 cents and adds names, phone numbers, and postal addresses for an additional 2 cents.[128,129]

[123] https://web.archive.org/web/20180330221239/https://datafinder.com/products/email-recovery.

[124] https://www.datazapp.com/reverse-email-append.

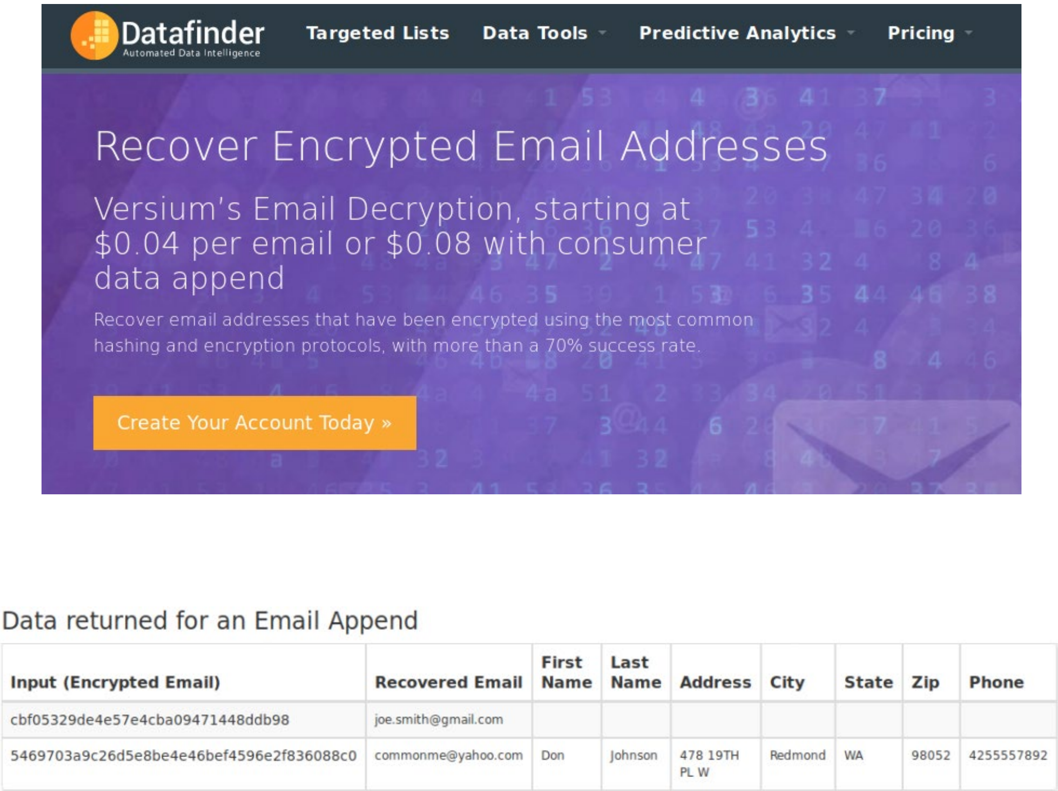
[125] https://www.datazapp.com/phone-append.

[126] Datazapp | How to Reverse Phone Append https://www.youtube.com/watch?v=B0tH4-9GMVI.

[127] Datazapp | How to Reverse Email Append + SHA & MD5 https://www.youtube.com/watch?v=ymesLfAXH0c.

[128] https://thedatagroup.com/reverse-email-append/.

[129] https://thedatagroup.com/reverse-phone-append/.



Recover Encrypted Email Addresses

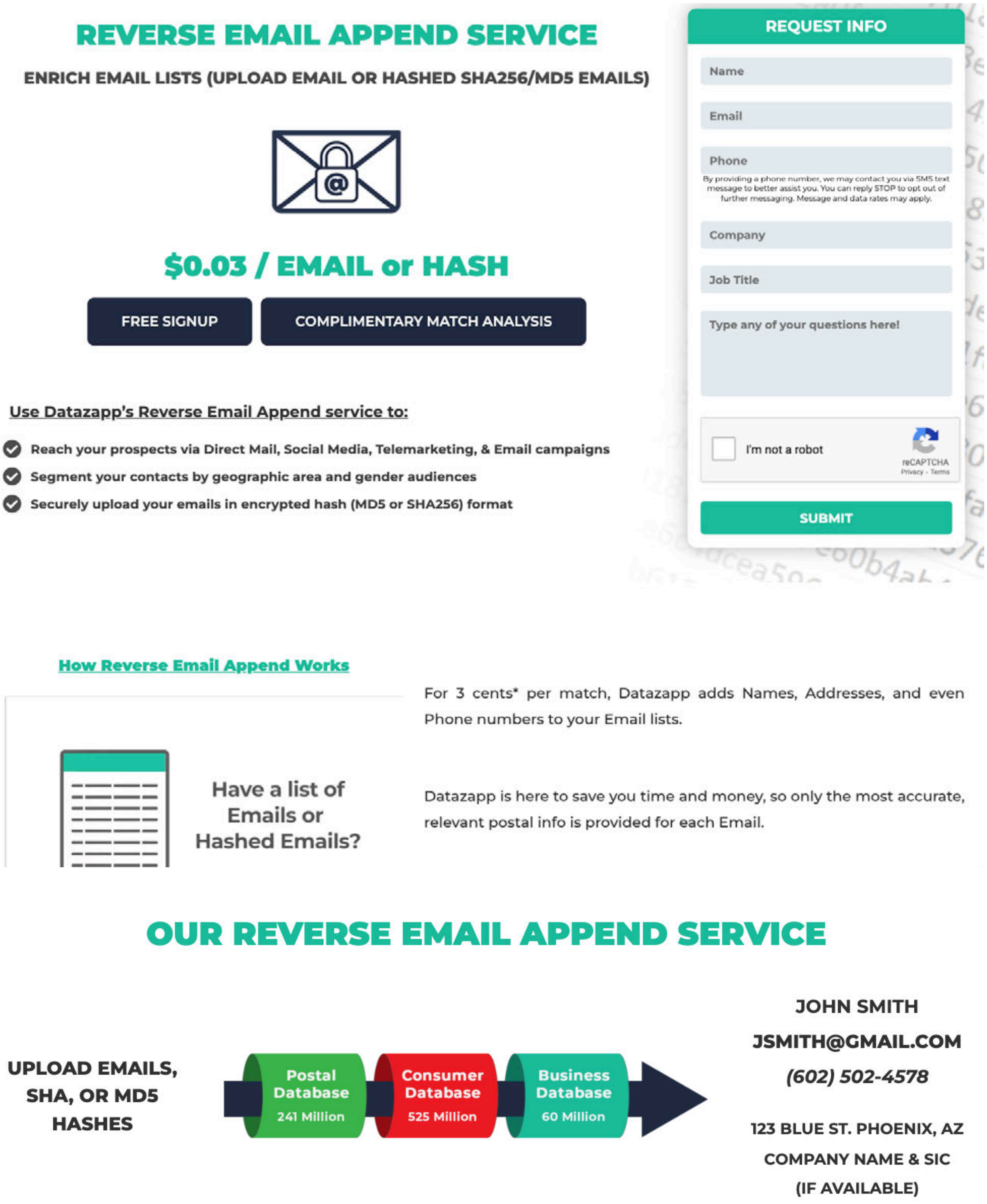
Versium's Email Decryption, starting at $0.04 per email or $0.08 with consumer data append

Recover email addresses that have been encrypted using the most common hashing and encryption protocols, with more than a 70% success rate.

Data returned for an Email Append

Input (Encrypted Email)	Recovered Email	First Name	Last Name	Address	City	State	Zip	Phone
cbf05329de4e57e4cba09471448ddb98	joe.smith@gmail.com							
5469703a9c26d5e8be4e46bef4596e2f836088c0	commonme@yahoo.com	Don	Johnson	478 19TH PL W	Redmond	WA	98052	4255557892

Figure 1: DataFinder service to reverse hashed email addresses



REVERSE EMAIL APPEND SERVICE

ENRICH EMAIL LISTS (UPLOAD EMAIL OR HASHED SHA256/MD5 EMAILS)

$0.03 / EMAIL or HASH

Use Datazapp's Reverse Email Append service to:

- Reach your prospects via Direct Mail, Social Media, Telemarketing, & Email campaigns
- Segment your contacts by geographic area and gender audiences
- Securely upload your emails in encrypted hash (MD5 or SHA256) format

REQUEST INFO

Name	
Email	
Phone	
Company	
Job Title	
Type any of your questions here!	

By providing a phone number, we may contact you via SMS text message to better assist you. You can reply STOP to opt out of further messaging. Message and data rates may apply.

☐ I'm not a robot

How Reverse Email Append Works

Have a list of Emails or Hashed Emails?

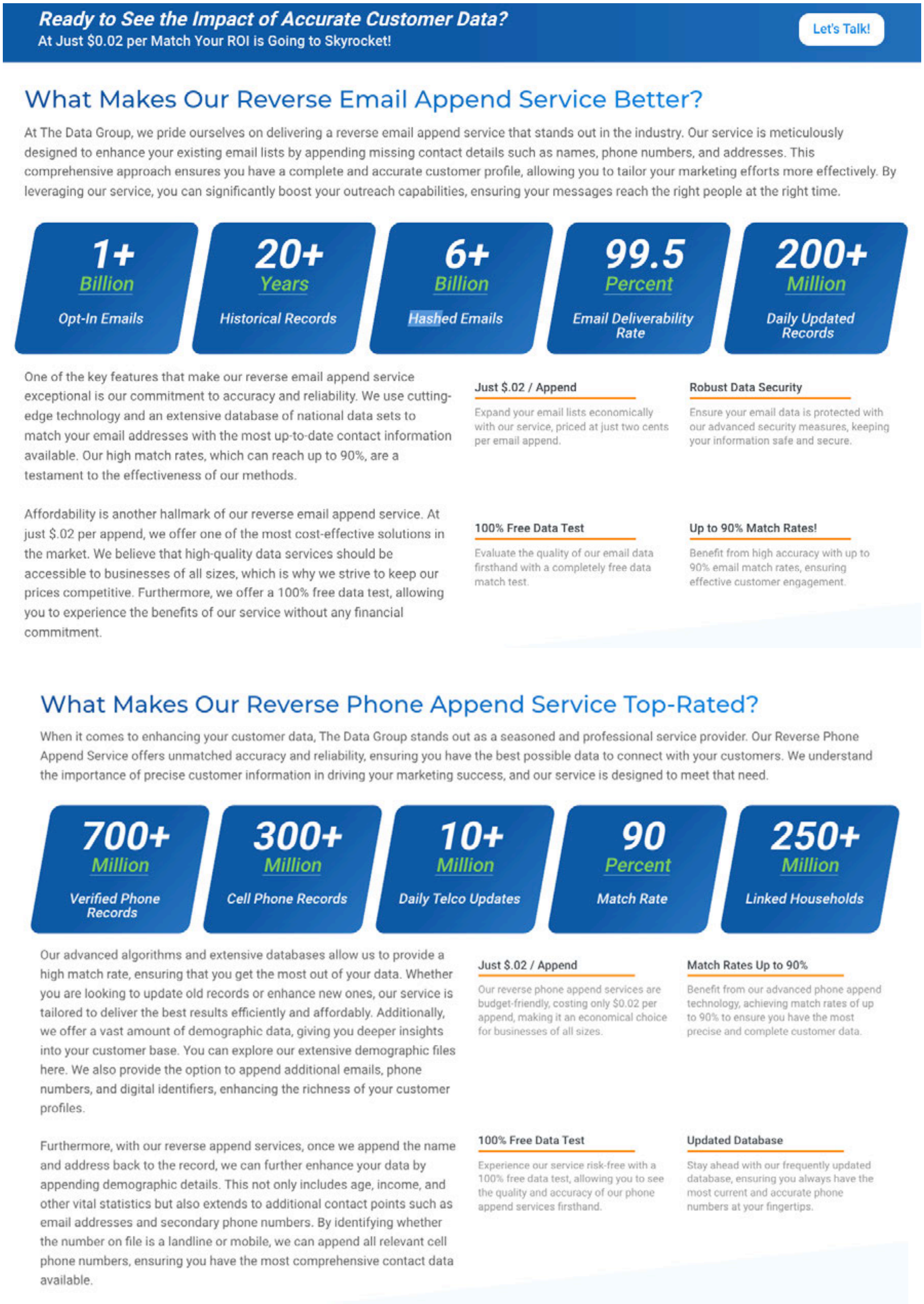
For 3 cents* per match, Datazapp adds Names, Addresses, and even Phone numbers to your Email lists.

Datazapp is here to save you time and money, so only the most accurate, relevant postal info is provided for each Email.

OUR REVERSE EMAIL APPEND SERVICE

UPLOAD EMAILS, SHA, OR MD5 HASHES
Postal Database 241 Million
Consumer Database 525 Million
Business Database 60 Million
JOHN SMITH
JSMITH@GMAIL.COM
(602) 502-4578
123 BLUE ST. PHOENIX, AZ
COMPANY NAME & SIC
(IF AVAILABLE)

Figure 2: DataZapp service to reverse hashed email addresses



Ready to See the Impact of Accurate Customer Data?
At Just $0.02 per Match Your ROI is Going to Skyrocket!

What Makes Our Reverse Email Append Service Better?

At The Data Group, we pride ourselves on delivering a reverse email append service that stands out in the industry. Our service is meticulously designed to enhance your existing email lists by appending missing contact details such as names, phone numbers, and addresses. This comprehensive approach ensures you have a complete and accurate customer profile, allowing you to tailor your marketing efforts more effectively. By leveraging our service, you can significantly boost your outreach capabilities, ensuring your messages reach the right people at the right time.

1+ Billion Opt-In Emails	20+ Years Historical Records	6+ Billion Hashed Emails	99.5 Percent Email Deliverability Rate	200+ Million Daily Updated Records

One of the key features that make our reverse email append service exceptional is our commitment to accuracy and reliability. We use cutting-edge technology and an extensive database of national data sets to match your email addresses with the most up-to-date contact information available. Our high match rates, which can reach up to 90%, are a testament to the effectiveness of our methods.

Affordability is another hallmark of our reverse email append service. At just $.02 per append, we offer one of the most cost-effective solutions in the market. We believe that high-quality data services should be accessible to businesses of all sizes, which is why we strive to keep our prices competitive. Furthermore, we offer a 100% free data test, allowing you to experience the benefits of our service without any financial commitment.

Just $.02 / Append

Expand your email lists economically with our service, priced at just two cents per email append.

Robust Data Security

Ensure your email data is protected with our advanced security measures, keeping your information safe and secure.

100% Free Data Test

Evaluate the quality of our email data firsthand with a completely free data match test.

Up to 90% Match Rates!

Benefit from high accuracy with up to 90% email match rates, ensuring effective customer engagement.

What Makes Our Reverse Phone Append Service Top-Rated?

When it comes to enhancing your customer data, The Data Group stands out as a seasoned and professional service provider. Our Reverse Phone Append Service offers unmatched accuracy and reliability, ensuring you have the best possible data to connect with your customers. We understand the importance of precise customer information in driving your marketing success, and our service is designed to meet that need.

700+ Million Verified Phone Records	300+ Million Cell Phone Records	10+ Million Daily Telco Updates	90 Percent Match Rate	250+ Million Linked Households

Our advanced algorithms and extensive databases allow us to provide a high match rate, ensuring that you get the most out of your data. Whether you are looking to update old records or enhance new ones, our service is tailored to deliver the best results efficiently and affordably. Additionally, we offer a vast amount of demographic data, giving you deeper insights into your customer base. You can explore our extensive demographic files here. We also provide the option to append additional emails, phone numbers, and digital identifiers, enhancing the richness of your customer profiles.

Furthermore, with our reverse append services, once we append the name and address back to the record, we can further enhance your data by appending demographic details. This not only includes age, income, and other vital statistics but also extends to additional contact points such as email addresses and secondary phone numbers. By identifying whether the number on file is a landline or mobile, we can append all relevant cell phone numbers, ensuring you have the most comprehensive contact data available.

Just $.02 / Append

Our reverse phone append services are budget-friendly, costing only $0.02 per append, making it an economical choice for businesses of all sizes.

Match Rates Up to 90%

Benefit from our advanced phone append technology, achieving match rates of up to 90% to ensure you have the most precise and complete customer data.

100% Free Data Test

Experience our service risk-free with a 100% free data test, allowing you to see the quality and accuracy of our phone append services firsthand.

Updated Database

Stay ahead with our frequently updated database, ensuring you always have the most current and accurate phone numbers at your fingertips.

Figure 3: DataGroup service to reverse hashed email addresses and phone numbers

57. Hence, it is not surprising that the Federal Trade Commission (FTC) has been warning for more than a decade[130] and most recently that "**hashes aren't 'anonymous' and can still be used to identify users, and their misuse can lead to harm. Companies should not act or claim as if hashing personal information renders it anonymized.**"[131] Unless additional security measures such as salting[132] or encryption[133] are used, plain SHA-256 hashing of email addresses and phone numbers as done by TikTok does not provide meaningful privacy protection.

58. Given that TikTok uses the plain SHA-256 hashing, [redacted][135] Given additional time, [redacted] This means that hashed

[130] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2012/04/does-hashing-make-data-anonymous.

[131] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/no-hashing-still-doesnt-make-your-data-anonymous (emphasis in original); *see also id.* ("Companies often claim and act as if data that lacks clearly identifying information is anonymous, but data is only anonymous when it can never be associated back to a person.").

[132] https://developer.mozilla.org/en-US/docs/Glossary/Salt ("In cryptography, **salt** is random data added to a password before it is hashed. This makes it impossible for an attacker to derive passwords from their hashes using precomputed tables of passwords and the corresponding hashes.").

[133] https://developer.mozilla.org/en-US/docs/Glossary/HMAC ("A **cryptographic hash function**, also sometimes called a digest function, is a cryptographic primitive transforming a message of arbitrary size into a message of fixed size, called a digest. Cryptographic hash functions are used for authentication, digital signatures, and message authentication codes.").

[134] Appendix P.2 [redacted] Scripts are in Appendix P.1.

[135] Appendix I.2 [redacted] Scripts are in Appendix I.1.

emails and phone numbers are just as identifying as plaintext emails and phone numbers because they are reversable.

59. [redacted]

[136] Due to the time constraint posed by extracting and analyzing the two one-day sample data sets for this report, as further described in [redacted] I did not have time to search for plaintiff data in the produced raw data, which may contain yet additional data TikTok collected from plaintiffs.

[137] [redacted]

[138] [redacted]

[139] [redacted]

[140] Appendix D.2 Plaintiff data in produced processed unmatched.xlsx, [redacted] tab, rows 2 to 11. The script is in Appendix D.1.

[141] [redacted]

[142] [redacted]

[143] [redacted]

[144] [redacted]

[redacted]

60. [redacted]

[145] [redacted] (produced as WATTERS-GRIFFITHTT000611).

[146] user_type = 0.

[147] [redacted] (produced as WATTERS-GRIFFITHTT000611).

[148] [redacted] tab. The script is in Appendix D.1.

[redacted]

61. [redacted]

62. [redacted]

[149] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[153] Appendix D.2 Plaintiff data in produced processed unmatched.xlsx, [redacted] The script is in Appendix D.1.

[154] Appendix D.2 Plaintiff data in produced processed unmatched.xlsx, [redacted] The script is in Appendix D.1.

[redacted]

[redacted]

[redacted]

[155] Appendix D.2 Plaintiff data in produced processed unmatched.xlsx, [redacted] tab, row 2. Column A [redacted] The script is in Appendix D.1.

[156] Appendix D.2 Plaintiff data in produced processed unmatched.xlsx, [redacted] tab, rows 3 to 16. Column A [redacted] The script is in Appendix D.1.

[redacted]

[redacted]

B. [redacted]

64. Since the data produced by TikTok was limited to two one-day time periods, I supplement my analysis of TikTok's produced data by analyzing whether the browsing data of the plaintiffs [redacted]

[157] Appendix E.2 [redacted]
The script is in Appendix E.1.

[redacted]

65. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[158] Files: [redacted] (produced as GRIFFITHTT002116), [redacted] (produced as GRIFFITHTT002117) and Appendix R Plaintiff Internet Artifacts and History Data.

[159] Files: [redacted] (produced as SHIH-GRIFFITHTT000183), [redacted] (SHIH-GRIFFITHTT000184), [redacted] (SHIH-GRIFFITHTT000185), [redacted] (SHIH-GRIFFITHTT000186), and Appendix R Plaintiff Internet Artifacts and History Data.

[160] Files: [redacted]x” (produced as WATTERS-GRIFFITHTT000610), [redacted] (WATTERS-GRIFFITHTT000611), [redacted] (WATTERS-GRIFFITHTT000612) and Appendix R Plaintiff Internet Artifacts and History Data.

[161] Appendix H.2 Output_Categorization of Plaintiff Internet Artifact Data. The scripts are in Appendix H.1. Appendix R Plaintiff Internet Artifacts and History Data.

[162] Appendix E.2 Output_ List of unique website domains across both days (March 28 and May 21) combined. The script is in Appendix E.1.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

66. Using Interactive Advertising Bureau's (IAB's) standard content taxonomy,[163] I further classified a subset of these webpages[164] in the plaintiffs' browsing history where TikTok collects data. To this end, I used the classification API provided by Website Categorization API.[165] [redacted]

[163] https://iabtechlab.com/standards/content-taxonomy/.

[164] As mentioned earlier, due to the time constraint posed by extracting and analyzing the two one-day sample data sets for this report (see Section IX), I did not have time to analyze and classify plaintiffs' browsing history in their complete "Internet History Data."

[165] https://www.websitecategorizationapi.com/categories.php.

[166] [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

67. As I explain further in greater detail in Section VI of this report, there is ample scientific evidence that shows that browsing information—even seemingly benign—can be used to infer sensitive information such as home/work address, gender, age, marital status, educational background, occupation, religious, political, and sexual associations as well as personality traits. Thus, the aforementioned examples, do not capture the full extent of the sensitive information that TikTok can infer from the browsing history of the plaintiffs collected by TikTok.

VI. TIKTOK'S COLLECTION OF IDENTIFYING AND SENSITIVE DATA FROM NON-TIKTOK USERS

A. TikTok's Collection of Identifying Data from Non-TikTok Users

1. *Identification of non-TikTok users using* [redacted] *collected by TikTok*

68. I first investigate TikTok's collection of identifying data such [redacted] for non-TikTok users in the two one-day samples of the processed data produced by TikTok for March 28, 2024 and May 21, 2024.[167]

69. I searched for [redacted] in the unmatched Pixel subset of the data produced by TikTok.[168] The sheer numbers below demonstrate the widespread prevalence of TikTok's collection of identifying data such as [redacted]

a. [redacted].[170]

b. [redacted].[172]

[167] Due to the time constraint posed by extracting and analyzing the two one-day sample data sets for this report (see Section IX), I only had time to analyze the unmatched Pixel data, but not Events API and raw data. I would expect the numbers to be much higher when Events API data is included. Furthermore, this analysis was limited to processed data.

[168] [redacted] Script is in Appendix B_Unmatched Pixel Data Appendix.

[169] Appendix I.2 [redacted] The scripts are in Appendix I.1.

[170] Appendix F.2 [redacted] The scripts are in Appendix F.1.

[171] Appendix J.2 [redacted] The scripts are in Appendix J.1.

[172] Appendix G.2 [redacted] The scripts are in Appendix G.1.

70. It is important to note that the two one-day samples of the processed data produced by TikTok for March 28, 2024 and May 21, 2024 [redacted]

71. Recall from earlier that hashed [redacted] can be trivially reversed and are thus just as identifying as [redacted] To further demonstrate that merely hashing [redacted] does not provide meaningful privacy protection, I conducted the following analysis on the two one-day sample data sets:

a. [redacted]

[173] Appendix B Unmatched Pixel Data.

[174] Appendix P.2 [redacted]. Scripts are in Appendix P.1.

[redacted].[175]

b. [redacted]

72. I next calculate the probability[178] that TikTok collected [redacted] from non-TikTok users.[179]

a. The two one-day unmatched Pixel processed data sets show that the probability that TikTok collects [redacted][180]

b. For an average non-TikTok user for whom TikTok collects [redacted],[181,182] the probability that TikTok would collect [redacted]

[175] Appendix G.2 [redacted] The scripts are in Appendix G.1.

[176] Appendix I.2 [redacted] The scripts are in Appendix I.1.

[177] Appendix I.2 [redacted] The scripts are in Appendix I.1.

[178] Ross, S.M., 2014. Introduction to probability models. Academic press.

[179] Due to time constraint posed by extracting and analyzing the two one-day sample data sets for this report (see Section IX), this analysis was only done on unmatched Pixel data in the processed data set and only on [redacted]. If [redacted] are considered, the percentages would be even higher.

[180] Count of unmatched events with plaintext or hashed [redacted] / Count of all unmatched events = [redacted] Appendix K [redacted]

[181] Appendix L.2 [redacted] Scripts are in Appendix L.1. March 28 average [redacted]

[182] Extrapolating from one day, the average number of data points collected by TikTok for a non-TikTok user is [redacted]

c. Conservatively, for a non-TikTok user for whom TikTok collects [redacted] [redacted] the probability that TikTok would collect [redacted]

d. Even more conservatively, for a non-TikTok user for whom TikTok collects [redacted] [redacted] the probability that TikTok would collect [redacted]

2. *Entropy analysis of the identifiers collected by TikTok for non-TikTok users*

73. Even beyond [redacted] Below I first provide entropy analysis of [redacted] and then describe how they are used in the industry to link them to a person or household.

74. The scientific community uses entropy as a privacy metric to quantify the risk of identifiability.[183] Entropy is measured in terms of bits. If the number of entropy bits for a piece of

[183] Laperdrix, P., Bielova, N., Baudry, B. and Avoine, G., 2020. Browser fingerprinting: A survey. ACM Transactions on the Web (TWEB), 14(2), pp.1-33.

Eckersley, P., 2010. How unique is your web browser? In Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings 10 (pp. 1-18).

Andriamilanto, N., Allard, T., Le Guelvouit, G. and Garel, A., 2021. A large-scale empirical analysis of browser fingerprints properties for web authentication. ACM Transactions on the Web (TWEB), 16(1), pp.1-62.

Steinbrecher, S. and Köpsell, S., 2003, March. Modelling unlinkability. In International workshop on privacy enhancing technologies (pp. 32-47). Berlin, Heidelberg: Springer Berlin Heidelberg.; Diaz, C., Seys, S., Claessens, J. and Preneel, B., 2002, April. Towards measuring anonymity. In International Workshop on Privacy Enhancing Technologies (pp. 54-68).

Clauß, S. and Schiffner, S., 2006, November. Structuring anonymity metrics. In *Proceedings of the second ACM workshop on Digital identity management*.

data exceed a certain threshold depending on the population size, that data can be used to uniquely identify (also known as fingerprint) users.

75. The entropy metric is also used to quantify the risk of identifiability in industry. For example, the Privacy Sandbox project uses entropy to determine the Privacy Budget.[184] Chrome browser uses entropy to label APIs "that exposes data that folks on the internet find useful for fingerprinting." Specifically, in Chrome source code, "Attributes and methods marked as [HighEntropy] are known to be practically useful for identifying particular clients on the web today."[185] The privacy non-profit public interest group Electronic Frontier Foundation also uses entropy as a metric to assess identifiability of information.[186] The World Wide Web Consortium (W3C)'s Privacy Interest Group (PING) also uses entropy to assess and mitigate the risk of fingerprinting, which is defined as "the capability of a site to identify or re-identify a visiting user, user agent or device."[187]

76. The amount of entropy required to uniquely identify someone in a population of size N is $\log_2(N)$. Given that Earth's population is approximately 8 billion, the number of required bits is $\log_2$ (8 billion) = 32.897 ≈ 33 bits. Given that the number of Internet users on Earth is ≈ 4

Serjantov, A. and Danezis, G., 2002, April. Towards an information theoretic metric for anonymity. In International Workshop on Privacy Enhancing Technologies (pp. 41-53).

Deng, Y., Pang, J. and Wu, P., 2007. Measuring anonymity with relative entropy. In Formal Aspects in Security and Trust: Fourth International Workshop, FAST 2006, Hamilton, Ontario, Canada, August 26-27, 2006.

[184] Privacy Budget: Limit the amount of individual user data exposed to sites to prevent covert tracking. https://developer.chrome.com/en/docs/privacy-sandbox/privacy-budget/.

[185] https://chromium.googlesource.com/chromium/src/+/main/third_party/blink/renderer/bindings/IDLExtendedAttributes.md#HighEntropy_m_a_c.

[186] A Primer on Information Theory and Privacy https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy.

[187] Mitigating Browser Fingerprinting in Web Specifications https://www.w3.org/TR/fingerprinting-guidance/.

billion, the number of required entropy bits to uniquely identify a user or device on the Internet is log2(4 billion) = 31.897 ≈ 32 bits.[188]

77. The industry uses 32 bits of entropy as the identifiability threshold. As shown below, Google uses the 32 bits as the identifiability threshold[189] in calculating "Privacy Budget." As another example, as shown below, the FAQ page of Google's Privacy Budget project explains the use of the 32-bit entropy threshold for identifiability.[190] According to W3C Privacy Interest Group (PING), "30-some bits of entropy would be enough to uniquely identify every individual person."[191]

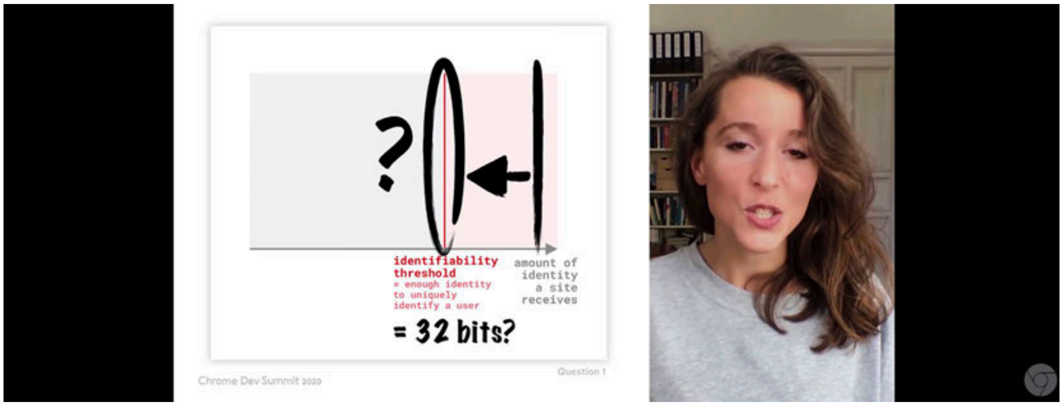


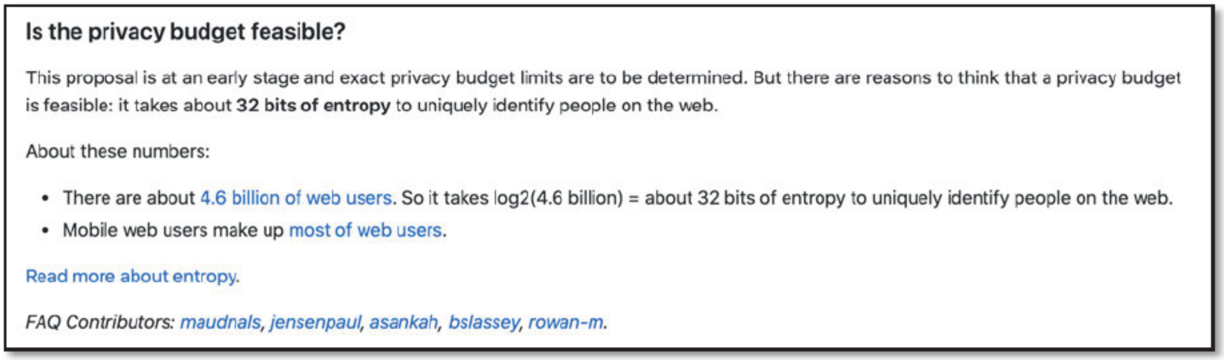
Figure 10: 32-bit identifiability threshold to uniquely identify a user

[188] https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy.

[189] Introducing the Privacy Budget https://www.youtube.com/watch?v=0STgfjSA6T8&t=423s.

[190] https://github.com/mikewest/privacy-budget/blob/4e5f78adde92bd622dafeceae78682fc0823c0eb/faq.md.

[191] https://w3c.github.io/fingerprinting-guidance ("Adding 1-bit of entropy is typically of less concern; 30-some bits of entropy would be enough to uniquely identify every individual person.").



Is the privacy budget feasible?

This proposal is at an early stage and exact privacy budget limits are to be determined. But there are reasons to think that a privacy budget is feasible: it takes about **32 bits of entropy** to uniquely identify people on the web.

About these numbers:

- There are about 4.6 billion of web users. So it takes log2(4.6 billion) = about 32 bits of entropy to uniquely identify people on the web.
- Mobile web users make up most of web users.

Read more about entropy.

FAQ Contributors: maudnals, jensenpaul, asankah, bslassey, rowan-m.

Figure 11: "It takes about 32 bits of entropy to uniquely identify people on the web"

78. I next use entropy to assess the identifiability of IP address, user agent, and cookies collected by TikTok. Specifically, I investigate if they exceed the 32-bit entropy threshold.[192]

79. First, IP address collected by TikTok by itself contains sufficiently identifying information to reach or exceed the 32-bit entropy threshold. There are two types of IP protocols: IPv4 and IPv6. The length of IPv4 address is 32 bits (i.e., approximately 4 billion possible IPv4 addresses). The new IPv6 protocol is now used by approximately half of the Internet users in the United States.[193] The length of an IPv6 address is 128 bits (i.e., approximately 340 trillion-trillion-trillion IP addresses possible IPv6 addresses). Thus, IP address, especially the newer IPv6 variant, can be sufficiently unique to meet or exceed the 32-bit threshold.

80. Note that IP addresses are sometimes reused or shared across users (e.g., using a mechanism called Network Address Translation [NAT] or Virtual Private Network [VPN]). However, the devices sharing an IP address are still distinguishable using the additional port number information that is always included alongside the IP address. Moreover, according to a

[192] A caveat to be aware of when calculating the "joint" entropy is that we should not simply sum up entropy of different pieces of data if they are dependent with each other. If different pieces of data are dependent, then the joint entropy could be lower than the simple sum of entropy of different pieces of data. Thomas M. Cover; Joy A. Thomas. Elements of Information Theory. Wiley (2006).

[193] https://www.google.com/intl/en/ipv6/statistics.html.

recent academic study, NAT and VPN correspond to a tiny fraction of all Internet traffic.[194] The study found that OpenVPN—the most widely used VPN implementation/protocol—is just 0.7% of all IPv4 traffic and 0.0% of all IPv6 traffic. The study found that NAT is just 0.5% of all IPv4 traffic and 0.0% of all IPv6 traffic.

81. While IP addresses may not always be static (i.e., they can change), peer-reviewed research[195] shows that the IP address by itself remains a serious threat to tracking despite the use of non-static IP addresses. Specifically, researchers showed that "87% of participants retain at least one IP address for more than a month."[196] For the study participants in the United States, the average IP address retention period was 18.93 days. Thus, IP address is a persistent identifier.

82. Second, the user agent collected by TikTok also contains a significant amount of entropy. There are approximately 10 bits of entropy in user agent according to one EFF study[197] and another AmIUnique study.[198] Thus, user agent when combined with other information such as IP address typically exceeds the 32-bit entropy identifiability threshold.

83. Third, companies typically store Universally Unique Identifiers (UUIDs) in cookies. "A UUID is 128 bits long, and can guarantee uniqueness across space and time."[199] TikTok also stores identifiers in the cookies that it collects.[200]

[194] Impact of Evolving Protocols and COVID-19 on Internet Traffic Shares, 2022.

[195] Mishra, V., Laperdrix, P., Vastel, A., Rudametkin, W., Rouvoy, R. and Lopatka, M., 2020, April. Don't count me out: On the relevance of IP address in the tracking ecosystem. In Proceedings of The Web Conference 2020 (pp. 808-815). https://dl.acm.org/doi/pdf/10.1145/3366423.3380161.

[196] *Id.*

[197] Peter Eckersley. How unique is your web browser? International Symposium on Privacy Enhancing Technologies Symposium, pages 1–18. Springer, 2010.

[198] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. IEEE Symposium on Security and Privacy, pages 878–894. 2016.

[199] A Universally Unique IDentifier (UUID) URN Namespace https://datatracker.ietf.org/doc/html/rfc4122.

[200] https://ads.tiktok.com/help/article/using-cookies-with-tiktok-pixel?lang=en.

84. In summary, IP address, user agent, and cookies collected by TikTok together easily exceed the 32-bit entropy identifiability threshold. Next, I describe how this identifying information is linked to persons or households by identity resolution services in the industry.

3. Identifiers collected by TikTok can be linked to a person or a household using identity resolution services

85. Identity resolution services match the identifiers such as IP address, user agent, and cookies to persons or households. Specifically, identity resolution services use *deterministic* or *probabilistic* matching to create an identity graph.[201,202] Deterministic matching is based on identifiers such as email address, phone number, and cookies. Probabilistic matching (also known as fingerprinting[203]) uses other identifiers such as IP address and user agent using statistical and machine learning techniques.

86. Internet Advertising Bureau (IAB), an industry consortium of advertising companies of whom TikTok is a member,[204] explains:

> "An ID solution is a product or a service that can help identify a person and/or household across digital environments e.g. web browsers, mobile apps, Connected Television (CTV) or other devices with which consumers interact and consume media. The shifting identity landscape has forced a transformation of ID solutions, relying on other signals not affected by browser and platform changes and offering consumers more control around their privacy choices. ID solutions have evolved to

[201] IAB Tech Lab Identity Solutions Guidance https://iabtechlab.com/wp-content/uploads/2024/05/Identity-Solutions-Guidance-FINAL.pdf ("ID solutions that rely on deterministic attributes of a consumer to identify the consumer that are relatively permanent and associated with one person or household are called deterministic IDs. Some examples include email address, phone numbers, and home address." "Some typical examples of attributes that are used for probabilistic IDs are IP address, user agent, timestamps, device details or settings.").

[202] https://adage.com/article/neustar/solving-identity-resolution-crisis/315290.

[203] https://www.incrmntal.com/resources/demystifying-problematic-measurement-deterministic-fingerprinting-and-probabilistic.

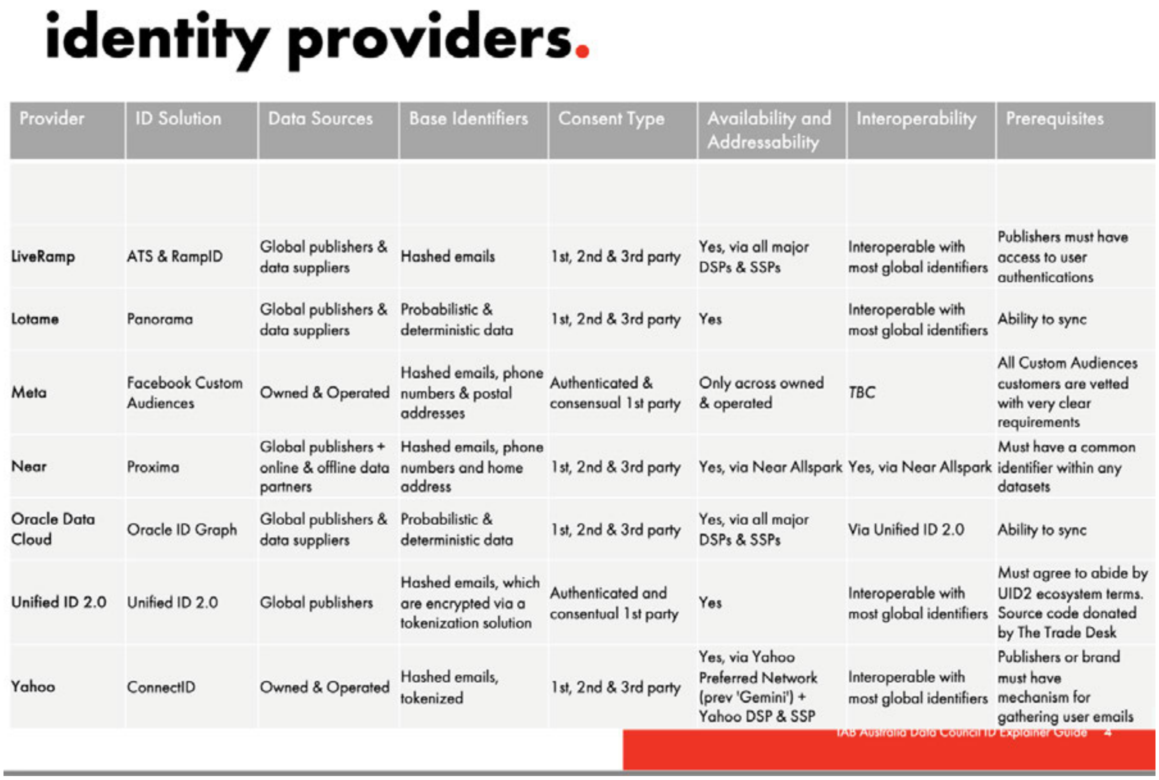
[204] https://www.iab.com/member-directory/.

> offer future-proofed ways of user identification in the absence of 3P cookies and device IDs."[205]

87. IAB lists popular identity resolution services provided by the likes of Experian, Equifax, TransUnion, Oracle, Criteo, LiveRamp, and Lotame.[206]

[205] IAB Tech Lab Identity Solutions Guidance https://iabtechlab.com/wp-content/uploads/2024/05/Identity-Solutions-Guidance-FINAL.pdf.

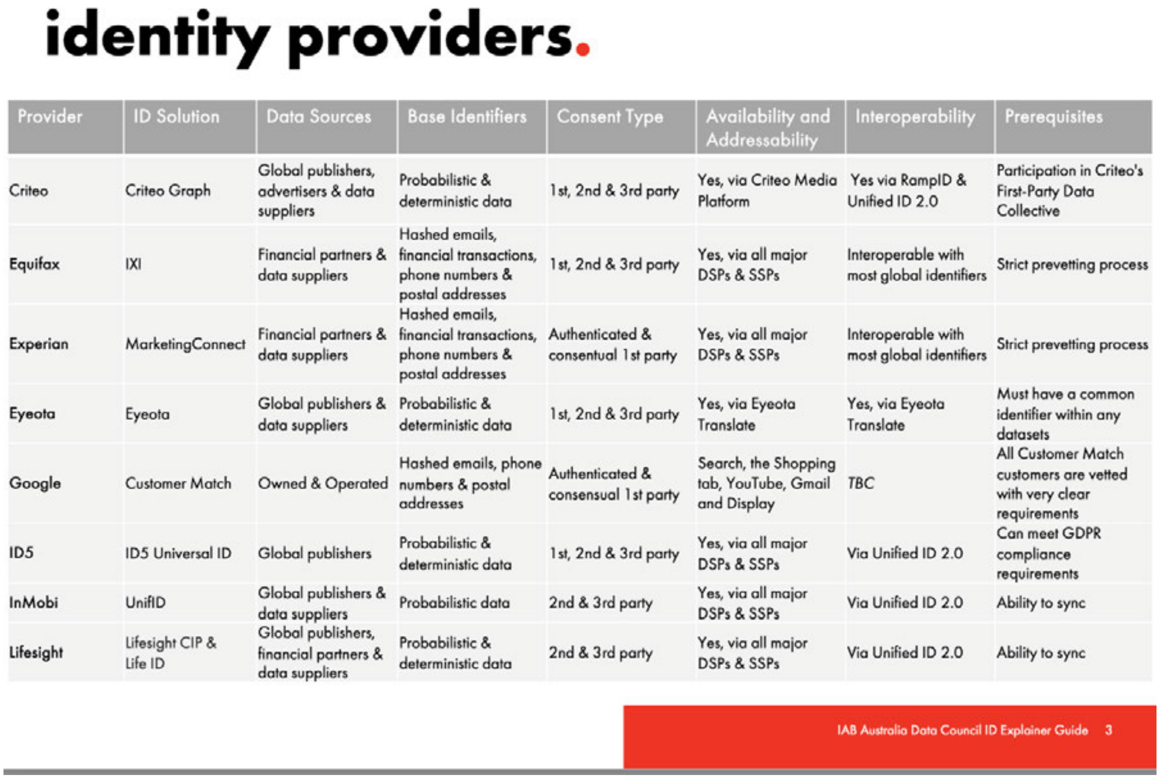
[206] https://m.iabaustralia.com.au/asset/284:id-matrixpdf.



identity providers.

Provider	ID Solution	Data Sources	Base Identifiers	Consent Type	Availability and Addressability	Interoperability	Prerequisites
LiveRamp	ATS & RampID	Global publishers & data suppliers	Hashed emails	1st, 2nd & 3rd party	Yes, via all major DSPs & SSPs	Interoperable with most global identifiers	Publishers must have access to user authentications
Lotame	Panorama	Global publishers & data suppliers	Probabilistic & deterministic data	1st, 2nd & 3rd party	Yes	Interoperable with most global identifiers	Ability to sync
Meta	Facebook Custom Audiences	Owned & Operated	Hashed emails, phone numbers & postal addresses	Authenticated & consensual 1st party	Only across owned & operated	TBC	All Custom Audiences customers are vetted with very clear requirements
Near	Proxima	Global publishers + online & offline data partners	Hashed emails, phone numbers and home address	1st, 2nd & 3rd party	Yes, via Near Allspark	Yes, via Near Allspark	Must have a common identifier within any datasets
Oracle Data Cloud	Oracle ID Graph	Global publishers & data suppliers	Probabilistic & deterministic data	1st, 2nd & 3rd party	Yes, via all major DSPs & SSPs	Via Unified ID 2.0	Ability to sync
Unified ID 2.0	Unified ID 2.0	Global publishers	Hashed emails, which are encrypted via a tokenization solution	Authenticated and consensual 1st party	Yes	Interoperable with most global identifiers	Must agree to abide by UID2 ecosystem terms. Source code donated by The Trade Desk
Yahoo	ConnectID	Owned & Operated	Hashed emails, tokenized	1st, 2nd & 3rd party	Yes, via Yahoo Preferred Network (prev 'Gemini') + Yahoo DSP & SSP	Interoperable with most global identifiers	Publishers or brand must have mechanism for gathering user emails

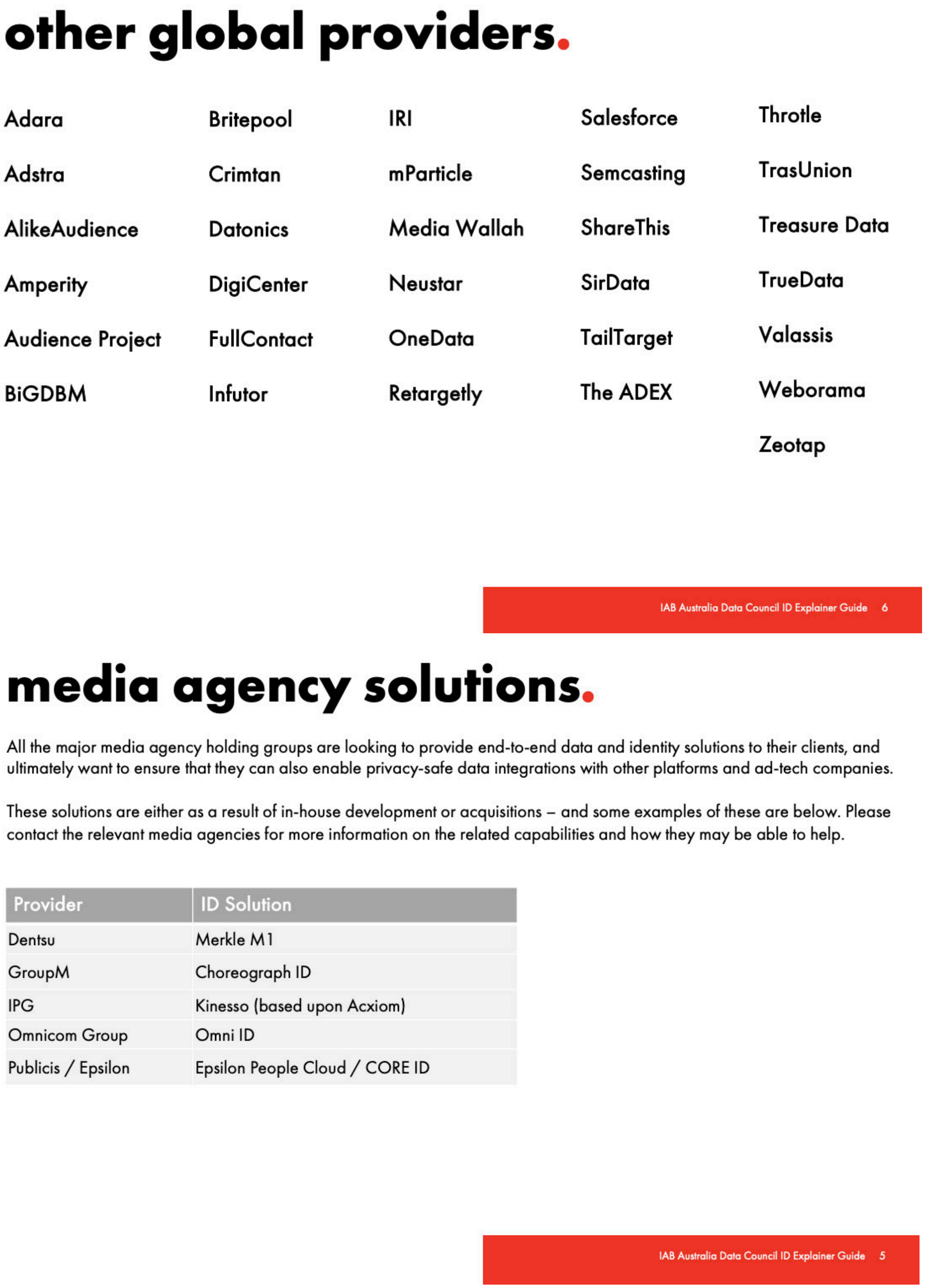
IAB Australia Data Council ID Explainer Guide 4



identity providers.

Provider	ID Solution	Data Sources	Base Identifiers	Consent Type	Availability and Addressability	Interoperability	Prerequisites
Criteo	Criteo Graph	Global publishers, advertisers & data suppliers	Probabilistic & deterministic data	1st, 2nd & 3rd party	Yes, via Criteo Media Platform	Yes via RampID & Unified ID 2.0	Participation in Criteo's First-Party Data Collective
Equifax	IXI	Financial partners & data suppliers	Hashed emails, financial transactions, phone numbers & postal addresses	1st, 2nd & 3rd party	Yes, via all major DSPs & SSPs	Interoperable with most global identifiers	Strict prevetting process
Experian	MarketingConnect	Financial partners & data suppliers	Hashed emails, financial transactions, phone numbers & postal addresses	Authenticated & consentual 1st party	Yes, via all major DSPs & SSPs	Interoperable with most global identifiers	Strict prevetting process
Eyeota	Eyeota	Global publishers & data suppliers	Probabilistic & deterministic data	1st, 2nd & 3rd party	Yes, via Eyeota Translate	Yes, via Eyeota Translate	Must have a common identifier within any datasets
Google	Customer Match	Owned & Operated	Hashed emails, phone numbers & postal addresses	Authenticated & consensual 1st party	Search, the Shopping tab, YouTube, Gmail and Display	TBC	All Customer Match customers are vetted with very clear requirements
ID5	ID5 Universal ID	Global publishers	Probabilistic & deterministic data	1st, 2nd & 3rd party	Yes, via all major DSPs & SSPs	Via Unified ID 2.0	Can meet GDPR compliance requirements
InMobi	UnifID	Global publishers & data suppliers	Probabilistic data	2nd & 3rd party	Yes, via all major DSPs & SSPs	Via Unified ID 2.0	Ability to sync
Lifesight	Lifesight CIP & Life ID	Global publishers, financial partners & data suppliers	Probabilistic & deterministic data	2nd & 3rd party	Yes, via all major DSPs & SSPs	Via Unified ID 2.0	Ability to sync

IAB Australia Data Council ID Explainer Guide 3



other global providers.

Adara
Adstra
AlikeAudience
Amperity
Audience Project
BiGDBM

Britepool
Crimtan
Datonics
DigiCenter
FullContact
Infutor

IRI
mParticle
Media Wallah
Neustar
OneData
Retargetly

Salesforce
Semcasting
ShareThis
SirData
TailTarget
The ADEX

Throtle
TrasUnion
Treasure Data
TrueData
Valassis
Weborama
Zeotap

IAB Australia Data Council ID Explainer Guide 6

media agency solutions.

All the major media agency holding groups are looking to provide end-to-end data and identity solutions to their clients, and ultimately want to ensure that they can also enable privacy-safe data integrations with other platforms and ad-tech companies.

These solutions are either as a result of in-house development or acquisitions – and some examples of these are below. Please contact the relevant media agencies for more information on the related capabilities and how they may be able to help.

Provider	ID Solution
Dentsu	Merkle M1
GroupM	Choreograph ID
IPG	Kinesso (based upon Acxiom)
Omnicom Group	Omni ID
Publicis / Epsilon	Epsilon People Cloud / CORE ID

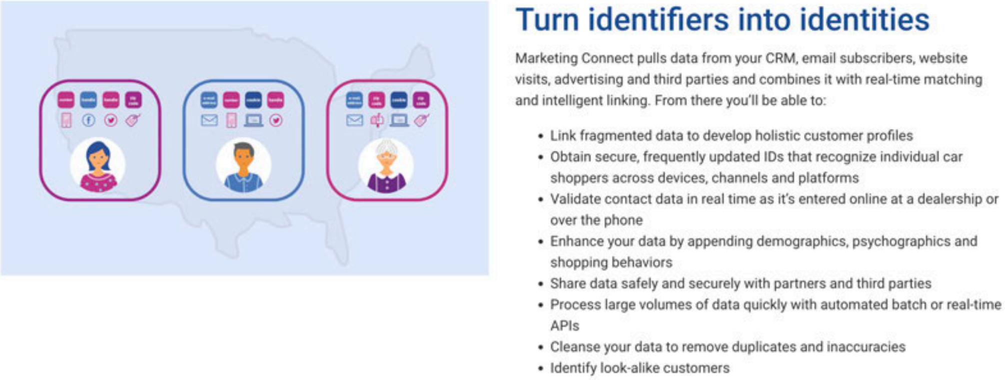
IAB Australia Data Council ID Explainer Guide 5

Figure 12: Identify services listed by the Internet Advertising Bureau (IAB)[207]

[207] https://m.iabaustralia.com.au/asset/284:id-matrixpdf.

88. Below are more detailed descriptions of some of these popular identity resolution services:

a. Experian: "Experian's identity resolution service, MarketingConnect, SM stitch together fragmented data such as names, addresses, emails, device IDs and cookies captured from multiple channels, platforms and devices to build holistic customer profiles. This single customer view ties to a persistent ID that makes it possible for brands to deliver a more personalized, omnichannel customer experience."[208]



Turn identifiers into identities

Marketing Connect pulls data from your CRM, email subscribers, website visits, advertising and third parties and combines it with real-time matching and intelligent linking. From there you'll be able to:

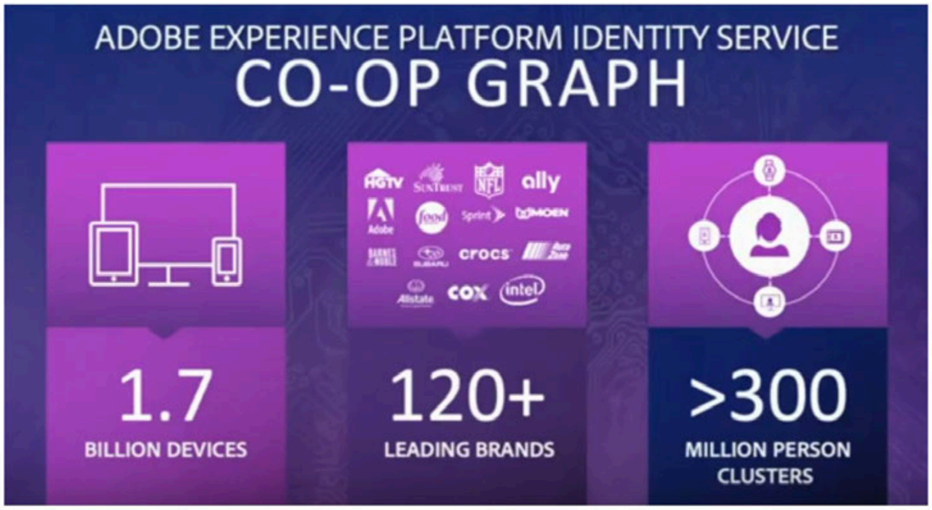
- Link fragmented data to develop holistic customer profiles
- Obtain secure, frequently updated IDs that recognize individual car shoppers across devices, channels and platforms
- Validate contact data in real time as it's entered online at a dealership or over the phone
- Enhance your data by appending demographics, psychographics and shopping behaviors
- Share data safely and securely with partners and third parties
- Process large volumes of data quickly with automated batch or real-time APIs
- Cleanse your data to remove duplicates and inaccuracies
- Identify look-alike customers

Figure 13: Experian's identity resolution service links identifiers such as IP address and cookies into identities

b. Adobe: "Adobe Experience Platform Identity Service accomplishes this by grouping device IDs into 'person clusters' that represent a pseudonymous person. Person clusters are identities based on deterministic data enriched with additional anonymous data associated with an individual through probabilistic matching."[209]

[208] https://www.experian.com/automotive/identity-resolution https://www.experian.com/blogs/marketing-forward/our-guide-to-identity-resolution/.

[209] https://blog.developer.adobe.com/adobe-experience-platforms-identity-service-how-to-solve-the-customer-identity-conundrum-f95e22d16ea9.



ADOBE EXPERIENCE PLATFORM IDENTITY SERVICE

CO-OP GRAPH

1.7 BILLION DEVICES

120+ LEADING BRANDS

>300 MILLION PERSON CLUSTERS

Figure 7: Adobe Experience Platform Co-op Graph provides one of the richest customer data sets for developing customer profiles on the market today.

Figure 14: Adobe's identity service links identifiers of 1.7 billion devices and 300 million persons using deterministic and probabilistic matching

c. The Trade Desk: "Identity graphs, like the ones used by our Identity Alliance, are essentially built in three steps. They're composed of a combination of deterministic identifiers ["Cookies, Mobile ad IDs, CTV IDs, Hashed emails, Unified ID 2.0 (UID2), RampID, IP addresses"] and probabilistic signals ["Wifi address, Time stamp, Geolocation, Browser attributes, Device attributes, User agent, Contextual data"] to cluster IDs at the household and individual level. The key to a great identity graph foundation is a sophisticated machine-learning model that can use the best data available to deliver scale and precision for targeting audiences across devices and channels."[210]

[210] https://www.thetradedesk.com/us/resource-desk/how-identity-graphs-are-built-the-present-and-the-future.

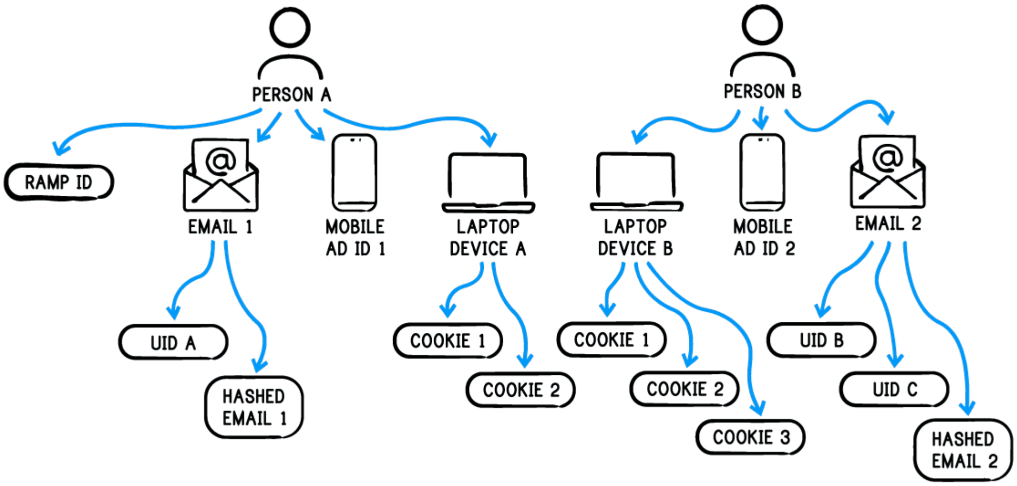
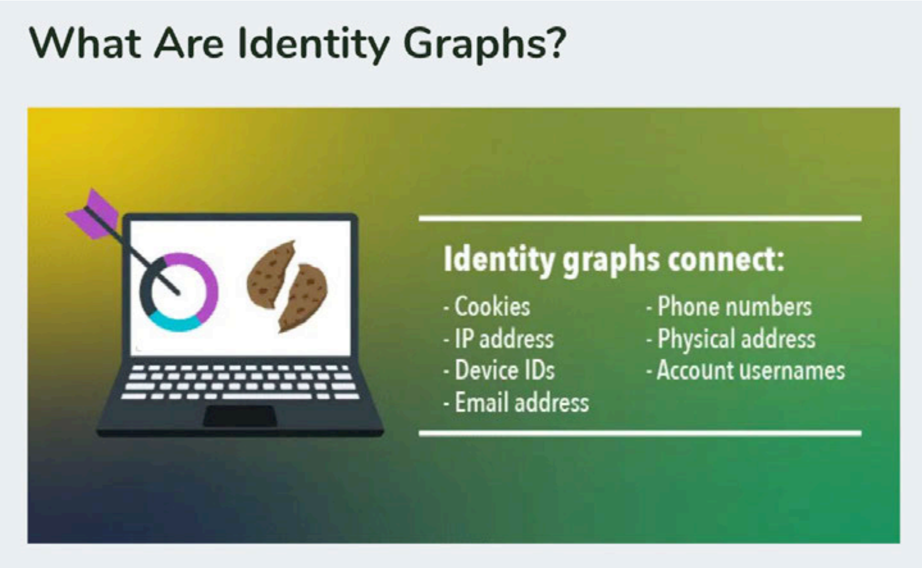


Figure 15: The Trade Desk's identity graph uses a combination of deterministic and probabilistic identifiers such as IP address and user agent to identify persons

d. Lotame: "Find your customers and prospects around the world with our patented graphing technologies, a mix of deterministic and probabilistic links. Panorama Graph connects and unifies consumer digital touch points across emails, cookies, and device IDs to offer a single view of a user." "Using a multitude of identifiers — web data, customer IDs, and hashed email — and both deterministic and machine learning approaches, our enriched identity solution makes it possible to reach and message the vast majority of consumers on connected devices. Additionally, Lotame Panorama ID supports and improves third-party tracking and eliminates dependence on the third-party cookie."[211,212]

[211] https://www.lotame.com/panorama-identity/.

[212] https://www.lotame.com/how-identity-graphs-benefit-a-connected-digital-advertising-ecosystem/.



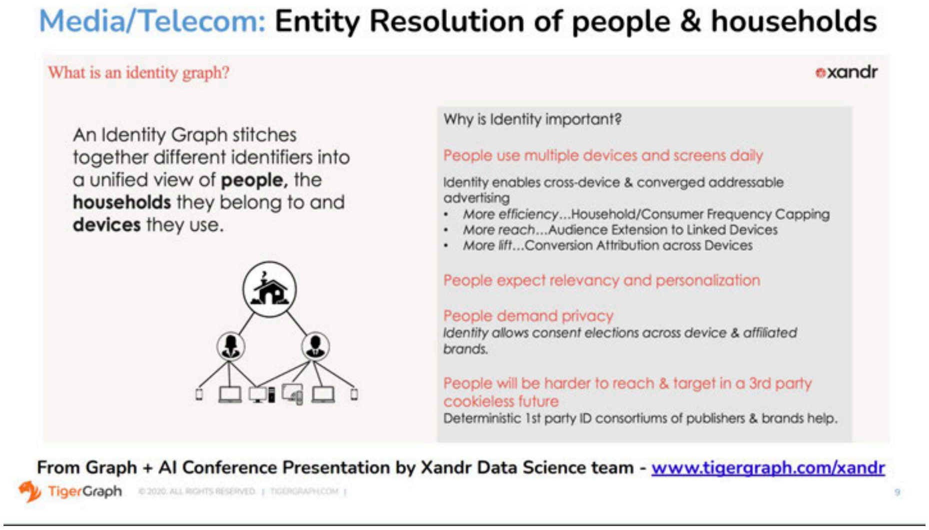
What Are Identity Graphs?

Identity graphs connect:

- Cookies
- IP address
- Device IDs
- Email address
- Phone numbers
- Physical address
- Account usernames

Figure 16: Lotame's identity graph connects IP address and cookies to email address, phone numbers, and physical addresses

e. AppNexus (now Xandr): "Xandr has built an identity graph using AT&T, WarnerMedia, Third-party and its own data, and leverage TigerGraph to perform deterministic and probabilistic entity resolution."[213]



Media/Telecom: Entity Resolution of people & households

What is an identity graph?

An Identity Graph stitches together different identifiers into a unified view of **people**, the **households** they belong to and **devices** they use.

Why is Identity important?

People use multiple devices and screens daily

Identity enables cross-device & converged addressable advertising

- *More efficiency*...Household/Consumer Frequency Capping
- *More reach*...Audience Extension to Linked Devices
- *More lift*...Conversion Attribution across Devices

People expect relevancy and personalization

People demand privacy

Identity allows consent elections across device & affiliated brands.

People will be harder to reach & target in a 3rd party cookieless future

Deterministic 1st party ID consortiums of publishers & brands help.

From Graph + AI Conference Presentation by Xandr Data Science team - www.tigergraph.com/xandr

Figure 17: Xandr's TigerGraph is an identity graph that "stiches together different identifiers into a unified view of people, the households they belong to and devices they use"

[213] https://info.tigergraph.com/hubfs/Misc./DSC%20Webinar%20Dec%2016%20-%20Knowledge%20Graph%20and%20Machine%20Learning%20-%203%20Key%20Business%20Needs%20One%20Platform%20.pdf.

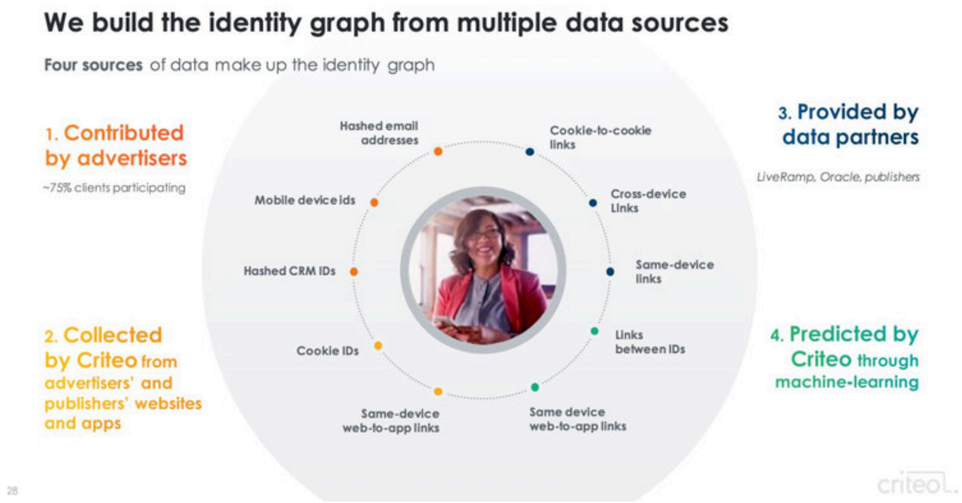
f. Criteo:[214] "The size, quality, persistence and focus of our graph make it a highly differentiated asset for identification.

i. **Large** 2B+ users, size comparable to Facebook's

ii. **High-Quality** Largely deterministic (matching 4 sources of identifier data with high level of certainty)

iii. **Persistent** 96% identities contain one or several persistent identifiers (hashed emails, logins, app IDs)

iv. **Focused** Focus on linking users to shopping data, unique outside Amazon."



We build the identity graph from multiple data sources

Four sources of data make up the identity graph

1. Contributed by advertisers

~75% clients participating

2. Collected by Criteo from advertisers' and publishers' websites and apps

3. Provided by data partners

LiveRamp, Oracle, publishers

4. Predicted by Criteo through machine-learning

Figure 18: Criteo uses hashed email addresses and cookies to build its identity graph

B. TikTok's Collection of Sensitive Data from Non-TikTok Users

1. *TikTok collects sensitive data from non-TikTok users*

89. I next investigate TikTok's collection [redacted] from non-TikTok users. To this end, I [redacted]

[214] https://criteo.investorroom.com/download/Criteo_Online_Identification_May2020.pdf.

[redacted]

90. [redacted]

[redacted] [redacted]

> [redacted]

[215] https://iabtechlab.com/standards/content-taxonomy/.

[216] https://www.websitecategorizationapi.com/categories.php.

[217] *See supra* at n.163 and accompanying text.

[218] Some URLs that were not classified by the Website Categorization API were excluded and additional URLs were randomly sampled until [redacted] were classified by the Website Categorization API.

[219] Appendix O.2 [redacted]. The scripts are in Appendix O.1.

[220] [redacted]

[221] [redacted]

[222] Appendix Q Script for Extracting Unique (full-string) URLs for March 28 and May 21.

[redacted]

[redacted]

c. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

1. [redacted]

[redacted]

m. [redacted]

[redacted]

92. I next calculate the probability that TikTok collected sensitive information from non-TikTok users.[223]

a. The two one-day unmatched Pixel processed data sets show that the probability that TikTok collects [redacted] from a non-TikTok user from a single event on a

[223] Due to time constraint posed by extracting and analyzing the two one-day sample data sets for this report (see Section IX), this analysis was done only on unmatched Pixel data in the processed data set and only on [redacted] If [redacted] are considered, the percentages would be even higher.

non-TikTok website is [redacted][224]

b. For an average non-TikTok user for whom TikTok collects 9[redacted], [225,226] the probability that TikTok would collect [redacted] from a non-TikTok user [redacted].

c. Conservatively, for a non-TikTok user for whom TikTok collects [redacted] [redacted], the probability that TikTok would collect [redacted] from that non-TikTok user [redacted]

d. Even more conservatively, for a non-TikTok user for whom TikTok collects [redacted] [redacted] the probability that TikTok would collect [redacted] from that non-TikTok user [redacted]

93. I have also analyzed [redacted] [redacted] [redacted]

[224] [redacted] Appendix K [redacted] Appendix B Unmatched Pixel Data.

[225] Appendix L.2 [redacted] Scripts are in Appendix L.1. [redacted]

[226] Extrapolating from one day, [redacted]

[227] Appendix M [redacted]

94. I next calculate the probability that TikTok collected [redacted] from non-TikTok users.[228]

a. The two one-day unmatched Pixel processed data sets show that the probability that TikTok collects [redacted] from a non-TikTok user from [redacted].[229]

b. For an average non-TikTok user for whom TikTok collects [redacted],[230,231] the probability that TikTok would collect [redacted] of a non-TikTok user [redacted]

c. Conservatively, for a non-TikTok user for whom TikTok collects just [redacted] the probability that TikTok would collect [redacted] of the non-TikTok user [redacted]

d. Even more conservatively, for a non-TikTok user for whom TikTok collects [redacted], the probability that TikTok would collect [redacted] of the non-TikTok user [redacted]

[228] Due to time constraint posed by extracting and analyzing the two one-day sample data sets for this report (see Section IX), this analysis was only done on unmatched Pixel data in the processed data set and only on [redacted] If [redacted] are considered, the percentages would be even higher.

[229] Appendix M [redacted]

[230] Appendix L.2 [redacted] Scripts are in Appendix L.1. [redacted]

[231] Extrapolating from one day, [redacted]

2. *Sensitive demographic and personality attributes can be inferred from seemingly benign browsing history*

95. Even beyond the sensitive browsing information collected by TikTok from non-TikTok users, there is ample scientific evidence that shows that browsing information—even seemingly benign—can be used to infer sensitive information such as home/work address, gender, age, marital status, educational background, occupation, religious, political, and sexual associations as well as personality traits.[232]

96. Based on a meta-review of 327 studies, Goel et al. concluded that demographic attributes such as age, sex, race, education, and income can be predicted from browsing history of users.[233] Hu et al. showed that age and gender can be used to can be predicted from the browsing history of users.[234] Murray and Durrell showed that sex, age, marital status, education, and whether or not one had children can be predicted from browsing history.[235] Lien et al. showed that a user's personality traits (e.g., honesty/humility, neuroticism, extraversion, agreeableness, conscientiousness, and openness) and demographic information (e.g., gender, age, relationship status) can be predicted from their browsing history even if the browsing history does not contain

[232] Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. PloS one, 13(11).

[233] Goel, S., Hofman, J. and Sirer, M., 2012. Who does what on the web: A large-scale study of browsing behavior. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 6, No. 1, pp. 130-137).

[234] Hu, J., Zeng, H.J., Li, H., Niu, C. and Chen, Z., 2007, May. Demographic prediction based on user's browsing behavior. In Proceedings of the 16th international conference on World Wide Web (pp. 151-160).

[235] Murray, D. and Durrell, K., 1999, August. Inferring demographic attributes of anonymous internet users. In International Workshop on Web Usage Analysis and User Profiling (pp. 7-20). Berlin, Heidelberg: Springer Berlin Heidelberg.

full-string URLs.[236] Park et al. showed that personality traits, demographics, and shopping interests can be inferred from browsing history.[237]

97. Research shows that browsing history is highly identifying. Olejnik et al.'s study of 368,284 browsing histories showed that 98% of the browsing histories of length 4 or more websites were unique.[238] Bird et al.'s study of 52,000 browsing histories showed that 99.65% of the browsing histories were unique. Even restricting visibility to the 100 most frequented domains resulted in 95% browsing histories being unique.[239]

C. Vignette of TikTok's Collection of Identifying and Sensitive Data

98. I illustrate the above points about TikTok's collection of identifying (Section VI-A) and sensitive (Section VI-B) data from non-TikTok users using the following example vignette.

99. [redacted]

[236] Lien, C.Y., Bai, G.J. and Chen, H.H., 2019, October. Visited websites may reveal users' demographic information and personality. In IEEE/WIC/ACM International Conference on Web Intelligence (pp. 248-252).

[237] Park, S., Matic, A., Garg, K. and Oliver, N., 2018. When simpler data does not imply less information: A study of user profiling scenarios with constrained view of mobile HTTP (S) traffic. ACM Transactions on the Web (TWEB), 12(2), pp.1-23.

[238] Olejnik, L., Castelluccia, C., & Janc, A. (2012). Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. In 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs).

[239] Bird, S., Segall, I., & Lopatka, M. (2020). Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories. In SOUPS@ USENIX Security Symposium.

[240] [redacted]

[241] Appendix S.1 Output_Vignette of TikTok's collection. The scripts are in Appendix S.2.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted]

[redacted]

[redacted]

[redacted]

[redacted] [redacted] [redacted] [redacted] [redacted] [redacted] [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

D. TikTok's Data Collection for the Plaintiffs Versus Non-TikTok Users At Large

100. The data TikTok collected from [redacted] is typical of the data TikTok collected from non-TikTok users at large. I provide two sets of analysis to this end.

101. [redacted]

102. Second, I analyze [redacted]

a. [redacted]

b. [redacted]

c. This analysis shows that TikTok's collection of data from the [redacted]

[242] Appendix R [redacted]. The script is in Appendix E.1.

VII. TIKTOK'S USAGE OF NON-TIKTOK USER DATA

103. The publicly stated purpose of TikTok's data collection is for advertising. TikTok explains that the TikTok Pixel is used to "Measure traffic on your website," "Measure ad campaign performance," and "Optimize your campaigns and find new customers."[243] TikTok further describes that TikTok Pixel "collects customer data and browsing behavior from your store to optimize your ad targeting experience," that it "tracks your ads' impact on your website," and that it "can help measure campaign performance and better define your ad's audience."[244]

104. [redacted]

105. [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[243] https://ads.tiktok.com/help/article/tiktok-pixel.

[244] https://ads.tiktok.com/help/article/data-sharing-tiktok-pixel-partners.

[245] TIKTOK-BG-000002930 at -932.

- [redacted]

[redacted]

[redacted]

107. [redacted]

108. I reserve the right to amend, modify, and supplement the opinions on TikTok's usage of non-TikTok user data should new information, [redacted], become available to me.

VIII. PRIVACY POLICY ANALYSIS

109. I used a script to search for the words "TikTok" and "ByteDance" across the privacy policies of websites that accounted for 54% of the total number of unmatched pixel events

[246] Amended Response to Interrogatory No. 8 (Apr. 16, 2024).

[247] TIKTOK-BG-002532985.

[248] TIKTOK-BG-002532957.

[249] TIKTOK-BG-002532973.

in the two one-day sample data sets produced by TikTok.[250] Across these, only 7.5% websites mentioned TikTok in any capacity (not necessarily about data collection). There was no mention of ByteDance, with the exception of one website that stated the affiliation between TikTok and ByteDance. Only 2.5% of the websites mentioned TikTok's data collection in some form. None of the websites describe the full extent of data collection by TikTok Pixel and Events API.

IX. DATA ANALYSIS SCHEDULE AND PROCESS TO DATE

110. In various sections of this report, I referred to the time constraint posed by extracting and analyzing the two one-day sample data sets for this report and noted more comprehensive analysis that I could and would undertake with more time (an estimated two to four additional weeks). This section describes the timeline and schedule of the data analysis that was undertaken to date.

111. TikTok started to produce the May 21, 2024 sample data on July 19, 2024 and the March 28, 2024 sample data on August 1, 2024. [redacted] I and a team of consultants working under my supervision were able to download the full data set for the two days by August 7, 2024.

112. My team and I began analysis of the two sets of data immediately, and I have been diligently analyzing the data since. By August 15, 2024, I completed a preliminary review of the produced data, including a comparison of the data fields in the raw, processed, and aggregated data

[250] Appendix N.2 Output_Privacy Policy. To perform this analysis, I wrote a script to visit top three results from Google Search for that website's privacy policy and subsequently searched (case insensitive) the text on the resulting webpages for "TikTok" and "ByteDance."

in relation to TikTok's prior sample data production, as well as a statistical analysis of the cookies and the other identifiers in the processed data. Once I understood the data structure, I started to develop a data analysis plan. This sometimes necessitated a closer look at subsets of data. Given the volume of the produced data, each data examination required scripts to be written, tested, and run. Even simple data extraction can take several hours on a powerful cloud computation infrastructure.

113. By August 28, 2024, I formulated several analyses tracks. By around September 6, 2024, my team began to implement the various analyses tracks under my supervision and guidance.

114. As I discussed in various sections of this report, due to the limited time available, I was unable to carry out a complete analysis of Events API data, a comprehensive search for all unmatched data associated with cookies, and an analysis of the produced raw and aggregate data. I estimate that these analyses would take another two to four weeks to implement, validate, and report.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed this 20th day of September 2024, in Davis, California.

Zubair

Zubair, Shafiq, Ph.D.